Universidade de Aveiro Departamento de Matemática
2005

João Pedro Antunes Ferreira da Cruz

Algoritmos de Aproximação Estocástica com Valor do Passo Adaptativo

Universidade de Aveiro Departamento de Matemática
2005

João Pedro Antunes Ferreira da Cruz

Algoritmos de Aproximação Estocástica com Valor do Passo Adaptativo

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Matemática, realizada sob a orientação científica de Alexandre Plakhov, Professor Associado Convidado no Departamento de Matemática da Universidade de Aveiro.

o júri

presidente

Doutor José Rodrigues Ferreira da Rocha, Professor Catedrático da Universidade de Aveiro.

vogais

Doutor Luís Henrique Martins Borges de Almeida, Professor Catedrático do Instituto Superior Técnico da Universidade Técnica de Lisboa.

Doutor Pedro Nuno Ferreira Pinto Oliveira, Professor Associado com Agregação da Escola de Engenharia da Universidade do Minho.

Doutor António Manuel Pacheco Pires, Professor Associado do Instituto Superior Técnico da Universidade Técnica de Lisboa.

Doutor Manuel González Scotto, Professor Auxiliar da Universidade de Aveiro.

Doutor Alexandre Plakhov, Professor Associado Convidado da Universidade de Aveiro (Orientador).

agradecimentos

Quero expressar a minha gratidão a todos os que contribuiram para que a presente tese pudesse ser uma realidade.

Em particular, agradeço ao meu orientador Prof. Doutor Alexandre Plakhov o profissionalismo da sua orientação com um especial apreço pela motivação que me ajudou a manter a necessária abertura de espírito.

Expresso também o meu reconhecimento pela amizade manifestada em muitos momentos de incentivo e também pelos úteis ensinamentos ao Andrey Sarychev, Isabel Pereira, Andreia Hall, Paula Rama, Anabela Ramos e Iola Mara Ribeiro.

Agradeço aos pais Teresa e João Alberto, e à amiga Paula Pires o apoio que, neste período invulgar, me permitiu compreender traços da minha própria personalidade que são energia pura para alcançar sonhos.

Agradeço o apoio das seguintes instituições:

Agradeço o apoio financeiro da FCT e do FSE no âmbito do III Quadro Comunitário de Apoio (Medida 5 - Acção 5.3 - Formação Avançada de Docentes do Ensino Superior - Concurso nr. 2/5.3/PRODEP/2001).

Agradeço à equipa do Departamento de Matemática da Universidade de Aveiro o apoio geral e em particular o apoio informático indispensável à realização da tese. Ao Prof. Doutor António Caetano, Presidente do Conselho Directivo à data do início dos estudos, agradeço a especial brevidade com que problemas surgidos nas condições de trabalho foram resolvidos.

Agradeço aos Serviços de Documentação da Universidade de Aveiro o rápido apoio no fornecimento da bibliografia requerida.

para a Paula, Maria João e Nelito

resumo

Consideram-se os algoritmos iterativos de aproximação estocástica (AE) do zero de uma função dada quando o valor da função é perturbado aleatoriamente. A teoria da AE está bem desenvolvida para o caso em que o valor do passo do algoritmo é determinístico, dependendo apenas do número da iteração do algoritmo; em particular, foram elaborados algoritmos assimptoticamente optimais.

No entanto, em muitos problemas práticos abordados de forma heurística (em particular redes neuronais), verifica-se que são mais efectivos, num periodo transitório, os algoritmos cujo valor do passo é aleatório, sendo determinado através dos parâmetros correntes do algoritmo.

A tese concentra-se nos algoritmos onde o valor do passo aumenta caso os incrementos de aproximações consecutivas mantenham o sentido (indicando que o algoritmo está na "zona determinística"), e diminui no caso contrário (estando o algoritmo na "zona estocástica").

No algoritmo de Kesten, o passo mantém-se caso hajam dois incrementos com o mesmo sinal, e em caso oposto, o passo é decrementado. Na tese, o algoritmo é generalizado podendo o passo aumentar caso duas iterações ocorram no mesmo sentido. É demonstrada a convergência para o zero com probabilidade 1 para o caso de funções unidimensionais e multidimensionais com um único zero. É também demonstrada a normalidade assimptótica das estimativas.

Podem encontrar-se na literatura, algoritmos heurísticos de variação multiplicativa do passo para redes neuronais. Na tese, e para o caso de funções unidimensionais com vários zeros, é demonstrado com probabilidade 1 que estes algoritmos podem divergir ou convergir para uma vizinhança de um dos zeros. A adaptação do passo depende de dois parâmetros que determinam a precisão da vizinhança. Além desta regulação foi observado em simulações que para uma maior precisão é necessário um aumento do número de iterações.

São apresentados inúmeros exemplos numéricos que ilustram a vantagem dos novos algoritmos para o caso unidimensional. Para o caso multidimensional os algoritmos propostos não se mostraram efectivos.

abstract

We consider iterative algorithms of Stochastic Approximation (SA) of the zero of a function when the value of the function is randomly perturbated. SA theory is well established when the step-value is deterministic, depending only on the iteration number. In particular, asymptotical optimal algorithms were developed.

However, in many practical problems, the use of random step-value becomes more efective in a transitory stage, being determined by current iterations measures. This thesis concentrate on algorithms where step can increase if consecutive increments have the same sign (called `deterministic zone') and decrease otherwise (called `stochastic zone'). Algorithms of this kind were treated heuristically in several publications (particularly in neural networks literature).

In Kesten's algorithm, step is kept if two consecutive iterations have the same direction, and decreases otherwise. The thesis makes a generalization allowing the step to increase if successive increments have same sign and decrease otherwise. Almost sure convergence is demonstrated for the case of unidimensional and multidimensional functions with one zero. The asymptotic normality of estimations is also proved.

Algorithms with multiplicative step variation were tested for neural networks. In this thesis, and for the case of unidimensional functions of many zeros, is demonstrated that divergence or convergence to a neighborhood of some zero can occur depending on algorithm parameters. In case of convergence more precision of the neighborhood can be reached at the expense of more iterations.

Many numerical examples are presented showing the advantage of the new algorithms for the unidimensional case. For the multidimensional case of these algorithms no benefit was observed.

Conteúdo

Lista de Figuras

Capítulo 1

Introdução e Motivação

A presente tese é sobre algoritmos iterativos de aproximação ao zero x^*, ou a um dos zeros, de uma função φ em que a medição da função está sujeita a uma perturbação aleatória possivelmente não limitada. Estes métodos são designados por métodos de Aproximação Estocástica e o trabalho pioneiro foi feito por Robbins e Monroe [38] em 1951.

A estimação de x^* é em cada instante t uma variável aleatória x_t cujo valor depende da estimativa anterior x_{t-1}, da medida $y_t = \varphi(x_{t-1}) + \xi_t$, em que ξ_t é uma perturbação aleatória, e do passo γ_t. A equação que determina x_t é

$$x_t = x_{t-1} - \gamma_t y_t, \quad t = 1, 2, \ldots$$

em que diferentes algoritmos existem para a regulação do passo γ_t.

Dois novos algoritmos são introduzidos e ambicionam uma rápida aproximação a uma solução útil, mantendo a atenção pelo comportamento óptimo no limite. O primeiro algoritmo de pesquisa do único zero generaliza o algoritmo de Kesten [23, 33], cujo trabalho foi possibilitar ao passo γ_t manter-se ou mesmo aumentar para acelerar a proximidade ao zero x^*. O segundo algoritmo é motivado por técnicas heurísticas que variam o passo γ_t de forma multiplicativa [1, 2, 3, 17, 43, 44] com vista a uma rápida aproximação a uma solução útil e que tem aplicação em funções com vários zeros. Estas técnicas são resumidas na Secção 1.8.

O presente Capítulo descreve uma panorâmica sobre métodos e aplicações em Aproximação Estocástica e o caminho que levou aos novos algoritmos. O Capítulo 2 e 3 contêm o enunciado e a demonstração de resultados teóricos referentes a uma generalização do algoritmo de Kesten, e ao algoritmo de passo multiplicativo, respectivamente. Por fim, o Capítulo 4 apresenta o estudo por simulação de casos.

1.1 Aplicações e metodologias relacionadas

A principal aplicação encontrada na bibliografia dos métodos de Aproximação Estocástica é a optimização de uma função, por exemplo o erro quadrático, cujo gradiente φ é conhecido mas depende de parâmetros que mudam em cada iteração. Vamos referir quatro aplicações que podem ser analisadas como métodos de Aproximação Estocástica: estimação recursiva de parâmetros duma distribuição (área da estatística), filtros lineares no processamento de sinal (área da electrónica), redes neuronais (área da computação) e o ajuste manual com poucas iterações (trabalhos em laboratório).

Redes neuronais. O algoritmo de treino supervisionado das redes neuronais (panorâmica deste paradigma em [21]) quando treinadas *on-line*, isto é, cada iteração usa um padrão de treino e modifica nesse instante os parâmetros que definem a rede, é um exemplo duma classe de métodos que se enquadra na Aproximação Estocástica. Apesar da forte proximidade entre os conceitos e da imensa bibliografia e pontos de vista teóricos sobre redes neuronais, é invulgar a ocorrência de relações entre estes conceitos. O livro de Kushner e Yin (1997) [25] apresenta uma breve introdução sobre esta relação na Secção 2.2.

Estimação de parâmetros duma distribuição. Estimar recursivamente o parâmetro θ^* duma distribuição da qual temos acesso a exemplos independentes é também uma aplicação de algoritmos de Aproximação Estocástica. O enquadramento é o seguinte: temos acesso a uma sequência independente de variáveis aleatórias Z_t provenientes duma distribuição de densidade $f_\theta(z)$ de onde pretendemos estimar θ. O estimador a definir $\theta_t := \hat{\theta}(Z_1, \ldots, Z_t)$ deverá satisfazer uma função objectivo que pretendemos minimizar, podendo ser, por exemplo, o erro quadrático $\mathrm{E}(\theta_t - \theta^*)^2$. Dependendo do problema em causa, o estimador pode ser recursivo $\theta_t = \theta_{t-1} - \gamma_t y_t$, em que y_t depende do problema e inclui directa ou indirectamente o efeito aleatório de Z_t, i.e., $y_t = y(\theta_{t-1}, Z_t)$. Nevel'son e Has'minskii (1971) [31] abordam o tema de forma extensa.

Filtro linear em controlo de sinal. Zero móvel. Trata-se de um caso muito específico e usado em telecomunicações. O objectivo é determinar um vector que estima o parâmetros de um comportamento linear dum processo. Existe uma vasta bibliografia de onde referimos a Secção 3 em Kushner e Yin (1997) [25], Benveniste, Metivier e Priouret (1990) [4] e o artigo de Delyon e Juditsky (1995) [15].

Zero móvel, processamento de sinal. A variação do passo γ_t nos algoritmos apresentados tem aplicação natural em funções a optimizar quando o óptimo (ou zero) é uma função

do tempo: $x^*(t)$. Os trabalhos no âmbito do processamento de sinal para o caso linear estão já muito estudados e já existem publicações com a generalização de algoritmos particulares heurísticos [19, 29, 30]. Existem trabalhos para outras funções de zeros móveis considerando a presença de perturbações aleatórias de que são exemplos [22, 28, 39].

Ajuste manual com poucas iterações. A Aproximação Estocástica pode funcionar como método simples de ajuste para a obtenção duma solução satisfatória em poucas iterações e sem recurso a métodos estatísticos padrão eventualmente mais complexos, como refere Fabian [16, Sec. 1.13]. A regulação da dose ideal da quantidade dum químico foi o exemplo dado. Outro exemplo antigo é o disparo de um canhão, exemplo descrito em Delyon (2000) [12] que cita B. Bru em 1890: o ângulo de disparo dum canhão para atingir uma certa região era ajustado por um factor multiplicativo $1/n$, para a n-ésima tentativa. Trata-se de um processo recursivo que faz uso apenas da memória da posição actual e não das anteriores posições estando sujeito a perturbações várias.

A não necessidade de memorizar toda a informação de exemplos e a possibilidade de aceitar perturbações inerentes ao problema são características principais dos algoritmos de Aproximação Estocástica. Várias são as obras de carácter geral sobre o tema das quais citamos Benveniste et. al. (1987) [4] e Kushner e Yin (1997) [25].

Comentário 1 *Pela clareza do texto, optámos pela expressão 'perturbação aleatória' na medida da função. Contudo, salientamos que esta terminologia não é sempre a mais adequada a cada uso, aplicação ou concretização dos métodos investigados no presente trabalho, como se depreende dos exemplos citados atrás.*

São ainda apresentados dois paradigmas relacionados com Aproximação Estocástica:

Optimização sem recurso ao gradiente. Logo depois do trabalho de Robbins-Monroe, surgiu o algoritmo para estimar o mínimo duma função quando não se tem acesso ao gradiente e em que as medidas dessa função estão sujeitas a uma perturbação. Kiefer e Wolfowitz (1952) são os autores do trabalho [24]. O gradiente é estimado efectuando duas medidas em cada iteração. Os algoritmos a que esta tese se refere usam apenas a medida de φ em cada iteração.

GMM. O método GMM (*Generalized method of moments*), estudado originalmente por Hansen (1982) [20], usa uma amostra de dimensão finita, possivelmente grande, por forma a estimar um parâmetro teórico β_0, escalar ou vectorial, usando simultaneamente estimadores escolhidos adequadamente para o problema e que de alguma forma definem uma função a minimizar $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^r$, em que $\mathbb{R}^p$ é o espaço onde vive o processo estocástico $\{x_t\}$,

$\mathbb{R}^q$ é o espaço do parâmetro a estimar, e $r \geq q$. Uma condição fundamental para o método convergir *quase-certamente* para o correcto valor do parâmetro é

$$\mathrm{E}[f(x_1, \beta_0)] = 0,$$

em que x_1 é uma variável aleatória (v.a.). Se pudermos considerar o gradiente de f temos um problema análogo à Aproximação Estocástica, com a diferença que toda a amostra é usada em simultâneo.

1.2 Optimalidade Assimptótica

Esta secção é sobre a eficiência da convergência de x_t para x^*. A variância $\mathrm{E}(x_t - x^*)^2$ é uma medida da velocidade de aproximação de x_t a x^*.

A consequência da presença de uma perturbação aleatória é o baixo decrescimento assimptótico da variância: decrescimento de ordem t^{-1}, muito inferior ao Método da Bissecção ou da Corda Falsa, para problemas determinísticos, que podem ter decrescimento de ordem exponencial 2^{-t} (por exemplo, [8]). A forma como a variância decresce no tempo está sujeita ao limite inferior de Rao-Crámer (por exemplo, [31, Cap. 8] e [32]) e tem sido dedicada muita atenção à obtenção de algoritmos cujo decrescimento em variância seja o mais rápido possível. Nos algoritmos de Aproximação Estocástica conhecidos é frequente o comportamento $\mathrm{E}(x_t - x^*)^2 = \mathrm{O}(1/t)$ e neste caso a constante

$$\lim_{t\to\infty} \mathrm{E}\left[\sqrt{t}(x_t - x^*)\right]^2$$

indica-nos uma medida da velocidade de aproximação das soluções. Este limite depende, além do algoritmo, dos dados do problema como a derivada $\varphi'(x^*)$ e a variância das perturbações $\mathrm{Var}\,\xi$.

Referimos, por completude, que sob certas condições iniciais de algoritmos de Aproximação Estocástica a variância do algoritmo pode ser pior que $\mathrm{O}(1/t)$, por exemplo, $\mathrm{O}(\ln t/t)$ ou mesmo $\mathrm{O}(1/t^a)$ para $a < 1$.

Um algoritmo designa-se por *algoritmo optimal* quando é obtida a taxa de decrescimento mínima no limite.

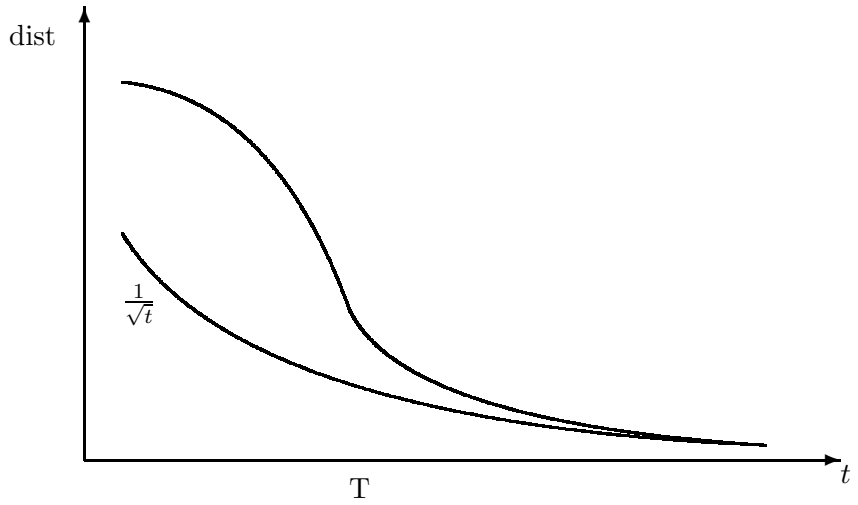


Figura 1.1: Tempos assimptóticos e transitórios.

1.3 Motivação: Rápida solução satisfatória

Para expor a motivação da tese introduzimos as noções qualitativas de tempos assimptóticos e tempos transitórios.

A análise feita na secção anterior foi feita para os tempos assimptóticos, ou seja, quando a trajectória esperada de $(x_t - x^*)^2$ tem o comportamento $\mathrm{O}(1/t)$, no melhor dos casos. Nesta situação, x_t oscila de forma aleatória em torno de x^*, reagindo às perturbações ξ_t, pois a magnitude de φ na vizinhança de x^* é pequena quando comparada com a magnitude das perturbações. Para uma pequena vizinhança de x^*, φ tem um comportamento quase linear sendo o declive $\alpha := \varphi'(x)$ um factor que influencia a velocidade da aproximação. O passo deve ser suficientemente pequeno para que não haja fuga para longe de x^*. A Figura 1.1 ilustra uma trajectória de x_t que para $t > T$ se comporta como $\mathrm{O}(1/\sqrt{t})$.

Ainda referente à mesma Figura, designamos a região $t \leq T$ por tempos transitórios sendo este período de duração caracterizada por vários factores: passo inicial γ_0 demasiado baixo ou demasiado elevado, ponto inicial x_0 numa região em que φ não pode ser comparado a uma recta com declive α, perturbações pequenas face à magnitude de $\varphi(x_0)$, e obviamente, o algoritmo escolhido.

Esta fase transitória pode ser muito prolongada no sentido do tempo útil disponível para se resolver determinado problema. A motivação da presente tese foi a criação de mecanismos de aceleração para encurtar os tempos transitórios, por forma a que se chegue mais rápido à fase assimptótica, na qual o comportamento é conhecido e eventualmente optimal.

É sobre este problema que a tese se debruça e são introduzidos dois novos algoritmos com o objectivo de melhorar a velocidade de convergência na fase não assimptótica. Antes, vamos apresentar algoritmos de Aproximação Estocástica que motivaram os propostos na tese.

1.4 O algoritmo padrão

O algoritmo padrão de Aproximação Estocástica é o seguinte:

Algoritmo 1 (Robbins-Monroe, 1951, [38]) *Consideremos uma função real φ com um único zero x^*, que verifica $(x - x^*)\varphi(x) > 0$ para $x \neq x^*$. A t-ésima aproximação ao zero x^* é dada por*

$$x_t = x_{t-1} - \gamma_t y_t, \tag{1.1}$$

$$y_t = \varphi(x_{t-1}) + \xi_t, \quad t = 1, 2, \ldots \tag{1.2}$$

em que ξ_t é uma perturbação aleatória, γ_t é uma sequência determinística que obedece a $\sum \gamma_t = \infty$ e $\sum_t \gamma_t^2 < \infty$, x_0 é uma condição inicial.

Duas considerações explicam a intuição do algoritmo:

1. Se $x_{t-1} < x^*$ o deslocamento médio $\mathrm{E}[x_t - x_{t-1}|x_{t-1}] = -\gamma_t\varphi(x_{t-1})$ é positivo e faz com que, em média, x_t de aproxime de x^*, e de modo análogo para o lado direito de x^*;

2. O passo γ_t no algoritmo padrão é uma sucessão determinística, positiva, que deve obedecer a $\sum_t \gamma_t = \infty$ e $\sum_t \gamma_t^2 < \infty$. A condição de divergência garante que o passo decresce de forma suficientemente lenta permitindo a x_t alcançar o zero da função; a não existência desta condição origina que x_t possa convergir mas não para o zero de φ. A segunda condição original é necessária à convergência do processo x_t; sem ela o processo diverge e, mesmo que o passo decresça, as aproximações x_t oscilariam em torno de x^* [4, Secção 2.2]. Esta condição pode ser generalizada para $\sum \gamma_t^\alpha < \infty$, $\alpha > 1$.

No trabalho original foi demonstrada a convergência em probabilidade e por Blum (1954) [6] a convergência *quase-certa*. A variante multidimensional foi obtida em [7] pelo memso autor.

O algoritmo com a escolha $\gamma_t = 1/(a \cdot t)$, $a > 0$ fixo, verifica as condições enunciadas para o passo e para este importante algoritmo de referência vamos enunciar propriedades assimptóticas probabilísticas considerando a condição $a < 2\varphi'(x^*)$.

Venter (1966) [47] demonstrou, para uma classe mais abrangente de algoritmos, que existe convergência em erro quadrático e, no caso do algoritmo e condição enunciadas acima, se tem $\mathrm{E}(x_t - x^*)^2 = \mathrm{O}(t^{-1})$. No livro de Nevel'son e Has'minskii (1971) [31] pode encontrar-se outra demonstração do mesmo resultado.

A distribuição da variável $\sqrt{t}(x_t - x^*)$, mediante as mesmas condições, converge em distribuição para uma Distribuição Normal [31, 42],

$$\sqrt{t}(x_t - x^*) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \frac{S^2}{a(2\alpha - a)},$$

em que $\alpha = \varphi'(x^*)$. São obtidas propriedades assimptóticas em [41, 10, 31]. A menor variância assimptótica é obtida tomando $a = \varphi'(x^*)$. No contexto da optimização, esta quantidade é a segunda derivada duma função f a minimizar tomando o valor no óptimo $f''(x^*)$.

Optimalidade assimptótica. No caso unidimensional $\varphi : \mathbb{R} \to \mathbb{R}$ a variância mínima possível, é

$$\mathrm{E}(x_t - x^*)^2 \geq \frac{S^2}{(\varphi'(x^*))^2} \cdot \frac{1}{t} \tag{1.3}$$

em que $S^2 = \mathrm{E}\xi^2$. Para o caso multidimensional $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ a variância mínima possível da norma das soluções é

$$\mathrm{E}(x_t - x^*)(x_t - x^*)^T \geq \varphi'(x^*)^{(-1)} \cdot S \cdot (\varphi'(x^*)^{(-1)})^T \cdot \frac{1}{t} \tag{1.4}$$

em que S é a matriz de covariâncias da perturbação aleatória $S = \mathrm{E}\xi\xi^T$ e $\varphi'(x^*)$ é a matriz Jacobiana no zero do campo de vectores φ, que no contexto de optimização é a Hessiana $\nabla^2 f(x^*)$ da função f multivariada a optimizar.

1.5 Medianização e optimalidade assimptótica

Consideremos um contexto em que é difícil obter ou lidar com a derivada ou matriz Jacobiana de φ em x^*, que é informação de 2ª ordem duma função a optimizar. Nestes casos, a optimalidade assimptótica pode ser atingida com *medianização*. A medianização consiste no cálculo, em paralelo, da média aritmética das estimativas x_t encontradas.

Algoritmo 2 (Ruppert, 1988, [40] e Polyak, 1990, [36]) *Consideremos uma função real φ com um único zero x^*, que verifica $(x - x^*)\varphi(x) > 0$ para $x \neq x^*$. A t-ésima aproximação ao zero x^* é dada por $\bar{x}_t$:*

$$x_t = x_{t-1} - \gamma_t y_t, \tag{1.5}$$

$$\bar{x}_t = \frac{1}{t}\sum_{i=1}^{t} x_i, \quad t = 1, 2, \ldots \tag{1.6}$$

em que a v.a. $\bar{x}_t$ para escolhas de γ_t que decrescem mais lentamente que $\gamma_t := 1/t$. Uma possível escolha do passo é $\gamma_t := \gamma \cdot t^{-a}$, onde $a < 1$ pertence a um intervalo que depende dos parâmetros do problema. x_0 é uma condição inicial. Assimptoticamente a variância mínima é atingida (por exemplo, [24, 13]).

Certos problemas em que haja restrições de tempo de execução dos algoritmos, possivelmente em detrimento da precisão, podem não ser satisfeitos com a técnica da medianização porque exemplos numéricos (ver [5, 45], entre outros) mostram que as primeiras iterações da medianização com o algoritmo padrão têm uma evolução mais lenta para o zero que outros métodos menos eficazes nos tempos assimptóticos.

1.6 Usando informação de segunda ordem

É oportuno agora referir que os algoritmos focados na tese usam apenas a informação da medida da função com o erro associado.

Existem métodos de segunda ordem, que estimam propriedades da função Hessiana e com isso obtêm uma mais rápida aproximação ao zero considerando o número de iterações. Esta vantagem pode ser perdida se o domínio da função em consideração tiver grande dimensão. Os trabalhos nesta área são vários dos quais referimos trabalhos aplicados à Aproximação Estocástica, como o algoritmo 'Kesten-Venter' de Tien (1977) [46], Spall (2000) [45] (escolha aleatória de coordenadas que entram na estimação da Hessiana), ou, especificamente no contexto das redes neuronais, o algoritmo de Levenberg-Marquardt, por exemplo Chen (2003) [9], LeCun et al. (1989) [27] em redes neuronais para reconhecimento de dígitos, ou ainda o estudo de Almeida et al. (1998) [1] onde uma matriz de passos é ajustada em cada iteração no contexto de optimização estocástica.

1.7 Algoritmos acelerados

De seguida iremos dar uma vista geral dos algoritmos que têm como objectivo acelerar a proximidade ao zero x^* na fase transitória. Todos se referem a funções com um único zero, unidimensionais ou multidimensionais.

É oportuno referir que usando a ideia de medidas extra em cada iteração, referida na Secção 1.1, relatamos a existência de trabalhos por forma a melhorar a velocidade nos tempos transitórios e também assimptoticamente. Por exemplo, o algoritmo 'Kesten-Venter' estudado

por Tien (1977) [46], garante a eficiência assimptótica com recurso a duas medições por iteração. Com alguma semelhança, Spall (2000) [45] introduz algoritmos que aproximam a matriz Hessiana e com isso pretendem acelerar a convergência na zona assimptótica. Nesta tese escolhemos usar apenas uma medida da função perturbada em cada iteração.

Os próximos algoritmos partilham a ideia comum para a adaptação do passo que consiste em observar se há mudança de sinal em duas diferenças consecutivas $\Delta x_{t-1} = -\gamma_{t-1} y_{t-1}$ e $\Delta x_t = -\gamma_t y_t$. A mudança de sinal em dois deslocamentos consecutivos é equivalente a observar o sinal de $y_t y_{t-1}$.

Algoritmo 3 (Kesten, 1957, [23]) *Seja* $\{\gamma(s),\, s = 0, 1, \ldots\}$ *uma sequência decrescente que verifica* $\sum_s \gamma(s) = \infty$ *e* $\sum_s \gamma^2(s) < \infty$.

$$x_t = x_{t-1} - \gamma(s_t) y_t, \quad t = 1, 2, \ldots \tag{1.7}$$

$$s_t = s_{t-1} + \mathbb{I}(y_t y_{t-1} \leq 0), \quad t = 2, 3, \ldots . \tag{1.8}$$

Neste algoritmo, o passo $\gamma(s_t)$ mantém-se se x_t se deslocar consecutivamente no mesmo sentido, ou seja, o mesmo que verificar se $y_t y_{t-1} > 0$. Neste artigo, Kesten estuda uma alteração ao algoritmo de Robbins-Monroe que pode acelerar a convergência para o zero na fase inicial. A ideia é que mudanças de sinal $(x_t - x^*) - (x_{t-1} - x^*) = x_t - x_{t-1}$ podem indicar que $|x_t - x^*|$ é pequeno e poucas mudanças de sinal de $x_t - x_{t-1}$ podem indicar que ainda se está longe do zero e, nesse caso, o passo deve ser mantido. São apresentados em [41] resultados de normalidade assimptótica sob condições restritas.

Importa salientar que no caso $\gamma(s) = 1/(a \cdot s)$, $a > 0$, a constante $a \cdot 2pq$, onde $p = \mathrm{P}(\xi > 0)$ e $q = \mathrm{P}(\xi \leq 0)$, tem efeito na variância limite

$$\mathrm{E}[\sqrt{t}(x_t - x^*)]^2 \to \frac{S^2}{a \cdot 2pq(2\alpha - a \cdot 2pq)}. \tag{1.9}$$

Para obter a menor variância no limite deve ajustar-se o parâmetro a por forma a que $a2pq = \varphi'(x^*)$, tal como no Algoritmo 1.

Algoritmo 4 (Delyon-Judistky, 1993, [14]) *O trabalho é análogo ao Algoritmo 3 de Kesten para o caso multidimensional* $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, *em que o produto interno* $y_t^T y_{t-1} \leq 0$ *é usado para determinar se o passo se deve manter ou diminuir.*

É apresentado o resultado de normalidade assimptótica e também uma versão com medianização para este algoritmo que permite alcançar a variância mínima no limite. São apresentados exemplos numéricos cuja formulação é usada na presente tese para ilustrar os novos

algoritmos. Alguma da estrutura da demostração deste trabalho serviu de base aos resultados obtidos para os novos algoritmos.

Algoritmo 5 (Plakhov-Almeida, 1999, [33]) *Seja* $\{\gamma(s),\, s = 0, 1, \ldots\}$ *uma sequência decrescente que verifica* $\sum_s \gamma(s) = \infty$ *e* $\sum_s \gamma^2(s) < \infty$.

$$x_t = x_{t-1} - \gamma(s_t) y_t, \quad t = 1, 2, \ldots \tag{1.10}$$

$$s_t = (s_{t-1} + \mathbb{I}(y_t y_{t-1} < 0) - \nu_t \cdot \mathbb{I}(y_t y_{t-1} \geq 0))^+, \quad t = 2, 3, \ldots \tag{1.11}$$

onde $\nu_t \in \{0, 1\}$ *com* $\mathrm{E}\nu_t = \nu$ *e* $a^+ = \max\{a, 0\}$.

O algoritmo apresentado neste trabalho contempla a possibilidade do contador de passo s_t diminuir quando o deslocamento ocorre no mesmo sentido causando o aumento do passo γ_t. Mas para assegurar o decrescimento do passo quando x_t está próximo de x^* a variável aleatória $\nu_t \in \{0, 1\}$ deverá ter uma distribuição tal que assegure que, quando x_t está próximo de x^*, o valor de s_t deverá aumentar garantindo a convergência de $\gamma_t \to 0$ e de $x_t \to x^*$. A convergência *quase-certa* do algoritmo foi demonstrada. O Algoritmo 3 de Kesten e este trabalho deram origem ao próximo algoritmo, mais geral, introduzido na presente tese.

1.8 Algoritmos propostos

São propostos dois algoritmos de Aproximação Estocástica com passo adaptativo.

Algoritmo 6 (Plakhov e Cruz, 2004, [35]) *Seja* $\gamma : \mathbb{R}_0^+ \to \mathbb{R}^+$ *uma função decrescente definida em* $[0, +\infty)$ *e que deve verificar* $\int_0^\infty \gamma(s)ds = \infty$ *e também* $\int_0^\infty \gamma^2(s) < \infty$.

$$x_t = x_{t-1} - \gamma(s_t) y_t, \quad t = 1, 2, \ldots \tag{1.12}$$

$$s_t = (s_{t-1} + \mathrm{u}(-y_t y_{t-1}))^+, \quad t = 2, 3, \ldots \tag{1.13}$$

em que $a^+ = \max\{a, 0\}$, u *é uma função sigmóide (crescente) e por isso o sinal* $-$ *no argumento da função* $\mathrm{u}(\cdot)$ *significa que pretendemos valores positivos de* $-y_t y_{t-1}$ *produzam aumento no contador de passo* s_t.

O Algoritmo 3 de Kesten e o Algoritmo 5 de Plakhov-Almeida usam a mudança de sinal entre y_{t-1} e y_t para determinar uma acção no contador do passo s_t. Além do sinal, este primeiro algoritmo introduzido na tese usa também a amplitude das alterações. A ideia é que se o sinal do deslocamento se mantem a amplitude de $-y_t \cdot y_{t-1}$ pode indicar um maior ou menor

afastamento ao zero x^* de φ. Se o afastamento é maior e o sinal igual então o passo deve aumentar e para isso o contador do passo deve decrescer mais. Se o sinal de y_t e y_{t-1} é oposto mas o valor absoluto $|y_t y_{t-1}|$ é grande então o passo deve decrescer rapidamente e para isso o contador do passo deve aumentar rapidamente. Este efeito de captar a dimensão de $-y_t y_{t-1}$ é codificado na escolha da função $\mathrm{u}(\cdot)$. As demonstrações dos resultados de convergência *quase-certa* e de normalidade assimptótica são introduzidos no Capítulo 2 da tese. Neste, é também apresentada a versão multidimensional do presente algoritmo. Assimptoticamente, é semelhante ao Algoritmo 1 de Robbins-Monroe em que a constante $\mathrm{E}_0 := \mathrm{E}[\mathrm{u}(-\xi_1\xi_2)]$, com ξ_1 e ξ_2 i.i.d., é análoga ao parâmetro a desse algoritmo.

Comentário 2 *Importa referir que as diferentes estratégias de adaptação do passo apresentam entre si um comportamento assimptótico do passo semelhante e que se resumem à formulação* $1/(\rho t)$*, em que* ρ *toma o valor* $\rho := a$ *no Algoritmo 1 de Robbins-Monroe,* $\rho := \mathrm{E}[\ \mathbb{I}(-\xi_{t-1}\xi_t)]$ *no Algoritmo 3 de Kesten,* $\rho := \mathrm{E}[\ \mathbb{I}(-\xi_{t-1}\xi_t)\nu_t]$ *no Algoritmo 5 de Plakhov e Almeida,* $\rho := \mathrm{E}[\mathrm{u}(-\xi_{t-1}\xi_t)]$ *no Algoritmo 6 introduzido na tese.*

O próximo algoritmo contempla funções com vários zeros convergindo para pontos da vizinhança de um dos zeros. A inspiração para o estudo deste algoritmo partiu de abordagens heurísticas com vista a acelerar a convergência nas redes neuronais. Nos trabalhos heurísticos de Silva e Almeida (1990) [44], Almeida et al (1998) [1], é usado o passo multiplicativo com o critério da mudança de sinal do gradiente para ajuste do passo entre iterações. Os trabalhos de Battiti (1989) [2] e Salomon e Hemmen (1996) [43] também usam passo multiplicativo sendo o critério o aumento ou diminuição do erro global.

Algoritmo 7 (Plakhov e Cruz, 2004, [34]) *Consideramos o problema de encontrar um dos zeros duma função* $\varphi : \mathbb{R} \to \mathbb{R}$*. O algoritmo tem a forma*

$$x_t = x_{t-1} - \gamma_t y_t, \quad t = 1, 2, \ldots, \tag{1.14}$$

onde o passo obedece à lei multiplicativa

$$\gamma_t = \begin{cases} \min\{u\gamma_{t-1}, \bar{\gamma}\} & \text{se } y_{t-1}y_t > 0, \\ d\gamma_{t-1} & \text{se } y_{t-1}y_t \leq 0, \end{cases} \quad t = 2, 3, \ldots \tag{1.15}$$

em que $0 < d < 1 < u$*,* $0 < \gamma_0, \gamma_1 < \bar{\gamma}$*, onde* $\bar{\gamma}$ *é uma constante positiva.*

Vale a pena sublinhar que este algoritmo usa a mudança de sinal na medida (perturbada) dos gradientes e é uma adaptação dos algoritmos definidos em [44, 1] ao problema de funções

unidimensionais com vários zeros. Na presente tese e no artigo citado [34] são estudadas de forma teórica as propriedades de convergência face a diferentes valores dos parâmetros u e d. Sob certas condições verifica-se que ocorre convergência se $ud < 1$ e divergência se $ud > 1$, deixando o caso $ud = 1$ em aberto. No caso de haver convergência, x_t converge para uma vizinhança de um dos zeros de φ sendo esta tão pequena quanto mais ud se aproxima, por valores inferiores, de 1. Uma maior precisão tem o custo de requerer mais iterações. Salvaguardamos, no entanto, uma ideia prática. Consideremos os pontos x_+^* e x_-^* ambos próximos dum zero de φ mas em que x_+^* está mais próximo desse zero. Ressalva-se que pode ocorrer que para constantes que verifiquem $u_1 d_1 < u_2 d_2 < 1$ se tenha que (u_1, d_1) produza uma trajectória para x_+^* e usando (u_2, d_2) se produza uma trajectória para x_-^* mais afastado do zero. Apenas a dimensão da vizinhança decresce uniformemente com a convergência de ud para 1.

O Capítulo 3 caracteriza com mais detalhe este algoritmo e demonstra a sua convergência ou divergência *quase-certa*.

Comentário 3 *A notação usada na tese associa o índice t de qualquer variável aleatória, como x_t, à σ-álgebra $= \sigma\{\xi_1, \ldots, \xi_t\}$. Na análise teórica dos algoritmos introduzidos, usamos o passo no instante $t-1$, atrasado face a x_t,*

$$x_t = x_{t-1} - \gamma_{t-1} y_t$$

e cuja modificação produz uma insignificante alteração de comportamento face ao algoritmo original. A vantagem desta técnica é uma análise teórica facilitada ao tornar o passo γ_{t-1} e a perturbação ξ_t, na medida $y_t = \varphi(x_{t-1}) + \xi_t$, duas variáveis aleatórias independentes.

Uma consequência desta técnica é a necessidade de duas condições iniciais s_0 e s_1 como se verá em detalhe nos próximos capítulos.

1.9 Resultados numéricos

No último Capítulo da tese apresentamos estudos numéricos dos métodos referidos para o caso duma função unidimensional, o caso de um campo de vectores originado pela função de Rosenbrock e ainda uma breve referência a uma rede neuronal para aprendizagem da função lógica ou-exclusivo.

O estudo foi realizado comparando o comportamento dos algoritmos: Robbins-Monroe

(RM) e Kesten (K), e ainda os algoritmos propostos, Kesten generalizado (Kg) e passo multiplicativo (Mul).

São constatadas as propriedades teóricas e traçadas conclusões práticas importantes sobre a aplicação dos algoritmos aos gerais problemas colocados.

Capítulo 2

Generalização do algoritmo de Kesten

Neste Capítulo apresentamos a demonstração de convergência *quase-certa* e resultados de normalidade assimptótica para o caso unidimensional e também para o caso multidimensional da generalização do algoritmo de Kesten introduzida na Secção 1.8.

2.1 Condições e Teoremas para o caso Unidimensional

Consideramos o problema de procura do único zero x^* de uma função φ real de variável real de acordo com o seguinte algoritmo de Aproximação Estocástica:

$$x_t = x_{t-1} - \gamma(s_{t-1})y_t, \quad t = 1, 2, \ldots \tag{2.1}$$

$$s_t = (s_{t-1} + \mathrm{u}(-y_t y_{t-1}))^+, \quad t = 2, 3, \ldots \tag{2.2}$$

onde

- $y_t = \varphi(x_{t-1}) + \xi_t$ é a $t-$ésima observação de φ perturbada com ruído aleatório ξ_t;
- $a^+ := \max\{a, 0\}$;
- u é uma função sigmóide;
- As variáveis aleatórias x_0, s_0, and s_1 são condições iniciais do algoritmo, possivelmente aleatórias;
- x_t é a $t-$ésima aproximação ao zero x^* de φ.

Na demonstração da convergência *quase-certa* consideramos que as seguintes condições são válidas:

Condições A1

1. $\{(s_0, x_0), \xi_1, \xi_2, \ldots\}$, tal como $\{s_1, \xi_1, \xi_2, \ldots\}$ são variáveis mutuamente independentes.
2. ξ_t são identicamente distribuídas, com média zero $\mathrm{E}\xi_t = 0$ e variância finita $S^2 = \mathrm{Var}\xi_t$.
3. Existe um Ω positivo tal que para cada intervalo $I \subset [-\Omega, \Omega]$, $\mathrm{P}(I) > 0$ (não ocorrem *falhas* em $[-\Omega, \Omega]$).
4. $\mathrm{E}|x_0| < \infty$.

Condições A2

1. $\gamma(s)$ é uma função positiva e decrescente definida em $[0, +\infty)$.
2. $\displaystyle\int_0^\infty \gamma(s)ds = \infty$.
3. $\displaystyle\int_0^\infty \gamma^2(s)ds < \infty$.

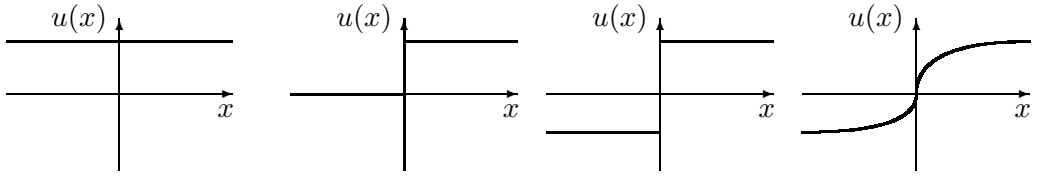
Condições A3

1. $\varphi \in \mathrm{C}^1(\mathbb{R})$, e $(x - x^*)\varphi(x) > 0$ quando $x \neq x^*$. Estas condições garantem o único zero x^* de φ.
2. Seja $M = \sup_x |\varphi'(x)|$ e $\beta_r = \inf_{|x-x^*|>r} \varphi^2(x)$. Para algum $R > 0$ verifica-se
$$\gamma(0) < \frac{2\beta_R}{(\beta_R + S^2)\, M}$$
(Esta condição limita superiormente o passo máximo $\gamma(0)$ e garante um infimo positivo para $\inf \varphi^2(x)$ quando $x\ to \pm\infty$).

Condições A4

1. u é uma função monótona crescente para a qual existem os limites finitos:
$$\mathrm{u}_+ = \lim_{x\to+\infty} \mathrm{u}(x) > 0 \text{ e } \mathrm{u}_- = \lim_{x\to-\infty} \mathrm{u}(x).$$
2. Denotamos $\mathrm{E}_\omega = \mathrm{E}[\mathrm{u}(X^{(\omega)})]$ onde
$$X^{(\omega)} = \inf_{\substack{|\varphi_1|\leq\omega \\ |\varphi_2|\leq\omega}} [-(\xi_1 + \varphi_1)(\xi_2 + \varphi_2)]$$
e denotamos o limite $\mathrm{E}_0 := \lim_{\omega\to 0^+}$. Requere-se que a constante E_0 seja positiva.

A Figura 2.1 exemplifica a função u.



(a) Caso do algoritmo de Robbins-Monroe.

(b) Caso do algoritmo de Kesten.

(c) Caso do algoritmo análogo ao de Plakhov-Almeida.

(d) Caso do algoritmo genérico.

Figura 2.1: Exemplos da função u.

Comentário 4 *Suponhamos que estamos a observar o processo (2.1), (2.2) começando de* $t_0 > 1$. *O novo processo com condições iniciais* x_{t_0}, s_{t_0}, s_{t_0+1} *e a sequência aleatória* $\xi_{t_0}, \xi_{t_0+1}, \ldots$ *também satisfazem a Condição A1.1 (Este comentário é usado no Lema 4).*

Comentário 5 *Se não existir ruído estocástico então* $S = 0$ *e a Condição A3.2 toma a forma* $M \cdot \gamma(0) < 2$. *De facto, esta condição é suficiente para* $|x_t - x^*|$ *decrescer de acordo com o algoritmo. (Ver Lema 24, pág. 62.)*

Comentário 6 *Se* u *ou a distribuição de* ξ_t *são contínuas, então* $\mathrm{E}_0 = \mathrm{E}[\mathrm{u}(-\xi_1\xi_2)]$. *Mais, se* u *é contínua e satisfaz* $\mathrm{u}(x) > -\mathrm{u}(-x)$ *quando* $x \neq 0$, *então A4.2 é válida para qualquer distribuição de* ξ_t *com variância não nula.*

Comentário 7 *De A1.3 segue que para qualquer intervalo* $I \subset [-\Omega, \Omega]$, $\mathrm{P}(I) \geq p(|I|)$, *onde* p *é uma função positiva mensurável.*

Teorema 1 (Plakhov e Cruz, 2004, [35]) *Sejam as Condições A1 a A4 válidas. Então, quase-certamente,* $\lim_{t\to\infty} x_t = x^*$.

As condições para se obter a normalidade assimptótica dos desvios $x_t - x^*$, são todas as condições de convergência A1 a A4 e ainda as Condições A4.3, A4.4 e A4.5.

Condição A4.3 $\mathrm{E}_0 < 2\alpha$ em que $\alpha := \varphi'(x^*)$.

Condição A4.4 φ admite decomposição de Taylor até à segunda ordem. Existe $D \geq 0$ tal que $|\varphi''(x)| \leq D$ para todo o x.

Condição A4.5 Assume-se a decomposição de Taylor da função $\mathrm{u}(x + \Delta x) = \mathrm{u}(x) + \mathrm{u}'(\theta)\Delta x$ para θ entre x e $x + \Delta x$.

Teorema 2 (Cruz, 2004) *Seja* x_t *definido por (2.1) e (2.2) para o qual se supõem verificadas as condições de convergência quase-certa de* $x_t \to x^*$. *Também se verificam as Condições A4.3, A4.4 e A4.5.*

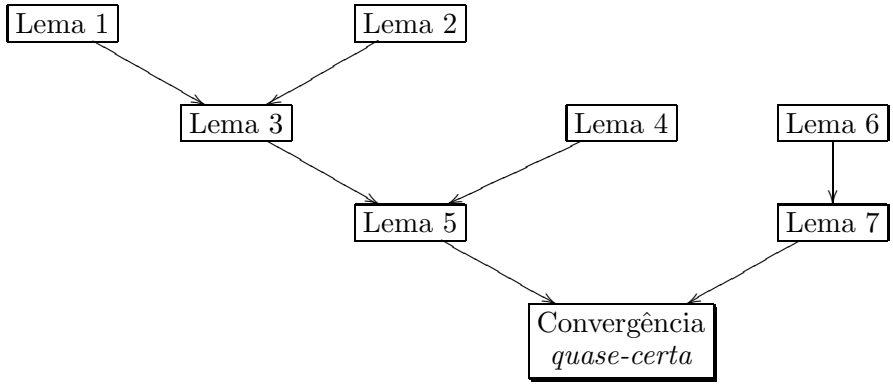


Figura 2.2: Lemas para a demonstração do Teorema 1.

Então, para $\gamma(s) = 1/s$ *(*$\xrightarrow{d}$ *designa convergência em distribuição),*

$$\sqrt{t}(x_t - x^*) \xrightarrow{d} \mathrm{N}\left(0, \frac{S^2}{\mathrm{E}_0(2\alpha - \mathrm{E}_0)}\right) . \tag{2.3}$$

2.2 Demonstração da Convergência *quase-certa*

Sem perda de generalidade supomos que $x^* = 0$. A demonstração segue os seguintes passos, que se encontram ilustrados na Figura 2.2,

- Para cada ϵ positivo, *quase-certamente* existe t tal que $|x_t| < \epsilon$ (Lema 3).
- Para cada ϵ e m positivos, *quase-certamente* existe t tal que $|x_t| < \epsilon$ e $\gamma_t < \gamma(m)$ (Lema 5).
- o 'contador do passo' s_t cresce, pelo menos linearmente, quando x_t permanece numa pequena vizinhança de x^*, facto que ocorre com probabilidade próxima de 1 (Lema 6).
- x_t permanece no intervalo $(-\epsilon, \epsilon)$ e s_t cresce pelo menos linearmente ($s_t > s_0 + \theta t - \eta$) para cada t com probabilidade positiva (Lema 7).

Fazemos agora uma descrição mais promenorizada das etapas das demonstrações. O Lema 3 mostra que, qualquer que seja $\epsilon > 0$, *quase-certamente* existe t tal que $|x_t| < \epsilon$. Este Lema baseia-se nos passos intermédios dos Lemas 1 e 2.

O Lema 1 baseia-se em duas vizinhanças de $x^* = 0$: uma tem raio R (valor obtido das condições que dependem de φ e variância S^2 das perturbações) e a outra vizinhança é

definida por um qualquer ϵ tão pequeno quanto se deseje. O Lema estabelece, que certamente, o trajecto x_t ou cai no intervalo $(-R, R)$ com um passo bastante elevado elevado ($\gamma_t > \gamma(m)$ onde m depende de ϵ) ou, caso contrário, cai num pequeno intervalo $(-\epsilon, \epsilon)$ porque se o passo for baixo, x_t pode alcançar a ϵ-vizinhança. Formalmente: para qualquer $\epsilon > 0$, *quase-certamente*, ou ocorre $|x_t| < \epsilon$, ou ocorre $|x_t| < R$ e $\gamma_t > \gamma(m)$ para um certo valor m que depende de ϵ.

A demonstração faz uso dum tempo-de-paragem definido com o evento complementar do acima mencionado. Mostra-se que a probabilidade do tempo-de-paragem ser finito (ou seja, não ocorrer o evento acima) é zero em limite, para todas as condições.

O Lema 2 estabelece um resultado para quando x_t e o passo γ_t partem duma situação inicial em que $|x_0| < R$ e $\gamma_0 > \gamma(m)$ (um qualquer m). Este Lema completa o resultado do Lema 1, pois sob estas condições iniciais, demonstra-se que existe probabilidade positiva, que depende apenas de m, de existir t tal que $|x_t| < \epsilon$.

A demonstração deste resultado é feita escolhendo pequenos intervalos adequados para cada ξ_t, por forma a usar a parte determinística do algoritmo.

Pelo Lema 1 e 2, verifica-se com probabilidade positiva δ que para qualquer condição inicial, existe t tal que $|x_t| < \epsilon$, em que δ depende apenas de ϵ (m depende de ϵ).

O que o Lema 3 traz de novo é que se este específico evento acima tem probabilidade positiva então a probabilidade dele ocorrer é 1.

A demonstração efectua-se por análise do evento "$|x_t|$ permanece fora da ϵ-vizinhança sempre" para todas as condições iniciais.

O Lema 5 mostra que existe t tal que $|x_t| < \epsilon$ e γ_t é tão pequeno quanto se deseja. Para este Lema é necessário o passo intermédio do Lema 4.

Para que o passo seja tão pequeno quando necessário requere-se que $y_t y_{t-1} < 0$, isto é, que o passo decresça consecutivamente, após x_t ter atingido uma ϵ-vizinhança dada. Para isso limitam-se os valores possíveis para os sucessivos ξ_t por forma a garantir $(-1)^t y_t > 0$, alternando o sinal de y_t, portanto. É ainda necessário garantir que x_t permanece na ϵ-vizinhança e por isso a amplitude de y_t também tem que ser controlada. Assim, o passo é tão pequeno quanto desejado e x_t permanece na vizinhança inicial, tudo isto com probabilidade positiva.

Se o Lema 4 garante que a probabilidade é positiva no evento "existe t tal que $|x_t|$ e γ_t são tão pequenos quando se deseje", então essa probabilidade é 1. A demonstração é análoga

à do Lema 3 e assim se tem a conclusão do Lema 5.

O Lema 6 constata uma propriedade sobre o comportamento do passo: assimptoticamente, o 'contador do passo' s_t cresce pelo menos linearmente, a uma taxa que depende apenas da função u(·) e da distribuição das perturbações.

A demonstração consiste em definir uma sequência $\tilde{s}_t > s_t$ à custa das variáveis ξ_t e ξ_{t-1}. Cada mudança de s_{t-1} para s_t faz-se à custa de y_t e y_{t-1}: daqui resulta que as mudanças em $\tilde{s}_t$ podem ser separadas em grupos de variáveis independentes e usar que soma de variáveis i.i.d. tem, assimptoticamente, distribuição normal.

O Lema 7 mostra que para $\epsilon_1 < \epsilon$, em que ϵ_1 é uma constante que depende de ϵ arbitrário, se $|x_0| < \epsilon_1$ a probabilidade para todo o t de x_t permanecer numa pequena vizinhança $(-\epsilon, \epsilon)$ e do 'contador do passo' s_t crescer linearmente, é positiva.

A conclusão da demonstração do Teorema é feita usando os Lemas 5 e 7 definindo o evento: 'x_t sai duma pequena vizinhança dada ou o passo γ_t é mais elevado que o pedido'. Caso este evento se realize, a probabilidade dele se voltar a realizar será sempre cada vez menor, e assim $|x_t|$ e γ_t são tão pequenos quanto se deseja, indefinidamente a partir de certo instante. Este é o resultado do Teorema.

Lema 1 *Para cada $\epsilon > 0$ existe $m = m(\epsilon)$ tal que quase-certamente ou ocorre (i) para alguns t, $|x_t| < \epsilon$, ou (ii) para alguns t, $|x_t| < R$ e $s_t \leq m$. (Relembramos que R está definido em A3.2.)*

Prova. Designamos por $f(x)$ a primitiva de $\varphi(x)$ satisfazendo $f(0) = 0$. Obviamente, $f(x) > 0$ quando $x \neq 0$. Fixamos $\epsilon > 0$ e definimos o tempo-de-paragem:

$$\tau = \tau(\epsilon, m) = \inf\{t : |x_t| < \epsilon \text{ ou } (|x_t| < R \text{ e } s_t \leq m)\}.$$

O objectivo é demonstrar que para algum m, $\mathrm{P}(\tau < \infty) = 1$.

Consideramos a sequência $\mathrm{E}_t = \mathrm{E}[f(x_t)\ \mathbb{I}(t < \tau)]$ e introduzimos a notação simplificada $f(x_t) = f_t$, $\mathbb{I}(t < \tau) = \mathbb{I}_t$, $f'(x_t) = f'_t$, $\gamma(s_t) = \gamma_t$. Usando que $\mathbb{I}_t \leq \mathbb{I}_{t-1}$,

$$\mathrm{E}_t - \mathrm{E}_{t-1} = \mathrm{E}[f_t I_t - f_{t-1} I_{t-1}] \leq \mathrm{E}[(f_t - f_{t-1})\ \mathbb{I}_{t-1}]. \tag{2.4}$$

Com a decomposição de Taylor

$$f_t = f(x_{t-1} - \gamma_{t-1} y_t) = f_{t-1} - f'_{t-1}\, \gamma_{t-1}\, y_t + \frac{1}{2} f''(x')\, \gamma_{t-1}^2\, y_t^2,$$

x' é um ponto entre x_t e x_{t-1}, substituindo $y_t = f'_{t-1} + \xi_t$ e relembrando que, de acordo com A3.2, $f''(x') \leq M$, obtemos

$$f_t - f_{t-1} \leq -\gamma_{t-1} f'_{t-1}\,(f'_{t-1} + \xi_t) + \frac{M}{2}\gamma_{t-1}^2 (f'_{t-1} + \xi_t)^2. \tag{2.5}$$

Usando (2.4) e (2.5) e tomando em conta que cada um dos valores γ_{t-1}, f'_{t-1}, $\mathbb{I}_{t-1}$ é determinado por x_{t-1} e s_{t-1}, e portanto mutuamente independentes de ξ_t (ver A1.1), ficamos com

$$\begin{aligned}
&\mathrm{E}_t - \mathrm{E}_{t-1} \leq \\
&\quad\leq\ \mathrm{E}[(-\gamma_{t-1} {f'_{t-1}}^2 - \gamma_{t-1} f'_{t-1}\xi_t + \frac{M}{2}\gamma_t^2 {f'_{t-1}}^2 + M\gamma_{t-1}^2 f'_{t-1}\xi_t + \frac{M}{2}\gamma_{t-1}^2\xi_{t-1}^2)\,\mathbb{I}_{t-1}] \\
&\quad=\ \mathrm{E}[(-{f'_{t-1}}^2 + \frac{M}{2}\gamma_t {f'_{t-1}}^2 + \frac{M}{2}\gamma_{t-1}S^2)\gamma_{t-1}\,\mathbb{I}_{t-1}].
\end{aligned} \tag{2.6}$$

Se $\mathbb{I}_{t-1} = 1$, ou ocorre 1) $|x_t| \geq R$, ou 2) $|x_t| \geq \epsilon$ e $s_t \geq m$. No caso 1) temos ${f'}^2_{t-1} > \beta_R$, e usando que $\gamma_{t-1} \leq \gamma(0)$ obtemos

$$\begin{aligned}
&-{f'_{t-1}}^2 + \frac{M}{2}\gamma_{t-1}{f'_{t-1}}^2 + \frac{M}{2}\gamma_{t-1}S^2 \leq \\
&\quad\leq\ -{f'_{t-1}}^2 + \frac{M}{2}\gamma(0){f'_{t-1}}^2 + \frac{M}{2}\gamma(0)S^2 \leq \\
&\quad\leq\ -\beta_R\,(1 - \frac{M}{2}\gamma(0)) + \frac{M}{2}\gamma(0)S^2 =: -c_0;
\end{aligned}$$

e de A3.2 segue que $c_0 > 0$.

No caso 2) temos $\gamma_{t-1} < \gamma(m)$. Relembrando que $\beta_\epsilon = \inf_{|x|\geq\epsilon}\varphi^2(x)$ temos

$$\begin{aligned}
&-{f'_{t-1}}^2 + \frac{M}{2}\gamma_{t-1}{f'_{t-1}}^2 + \frac{M}{2}\gamma_{t-1}S^2 \leq \\
&\quad\leq\ -\beta_\epsilon\,(1 - \frac{M}{2}\gamma(m)) + \frac{M}{2}\gamma(m)S^2 =: -c(\epsilon, m).
\end{aligned}$$

Escolhemos m tal que $c(\epsilon, m) > 0$ e denotamos $c = \inf\{c_0, c(\epsilon, m)\}$, então, em ambos os casos a expressão em parêntesis rectos no L.D. de (2.6) é menor que $-c\gamma_{t-1}\,\mathbb{I}_{t-1}$, então

$$\mathrm{E}_t - \mathrm{E}_{t-1} \leq -c \cdot \mathrm{E}[\gamma_{t-1}\,\mathbb{I}_{t-1}].$$

Usando que $s_t \leq s_0 + t\mathrm{u}_+$ e $\mathrm{E}\,\mathbb{I}_t = \mathrm{P}(t < \tau)$, obtemos

$$\mathrm{E}_t - \mathrm{E}_{t-1} \leq -c\,\gamma(s_0 + (t-1)\mathrm{u}_+)\,\mathrm{P}(t-1 < \tau),$$

e aplicando o argumento de indução e usando que $\mathrm{P}(j < \tau) \geq \mathrm{P}(t < \tau)$ quando $j < t$ obtemos

$$\mathrm{E}_t \leq \tilde{\mathrm{E}}_0 - c\sum_{j=0}^{t-1}\gamma(s_0 + j\mathrm{u}_+) \cdot \mathrm{P}(t < \tau),$$

onde $\tilde{\mathrm{E}}_0 := \mathrm{E}(f(x_0)\,\mathbb{I}(0<\tau)) < \infty$ pela Condição A1.4 (a constante E_0 definida na Condição A4 nada tem que ver com $\tilde{\mathrm{E}}_0$). A função f é positiva, por isso $\mathrm{E}_t \geq 0$, e daqui segue que

$$\mathrm{P}(t<\tau) < \frac{\tilde{\mathrm{E}}_0}{c\sum_{j=0}^{t-1}\gamma(s_0+j\mathrm{u}_+)}.$$

Comentário 8 *O majorante encontrado não depende do evento $\xi_1,\ldots,\xi_t$ permitindo que a conclusão do Lema seja válida para todos os eventos.*

Fazendo $t\to\infty$ e usando que, de acordo com A2.2, $\sum_{j=0}^{\infty}\gamma(s_0+j\mathrm{u}_+)=\infty$, concluímos que $P(\tau=\infty)=0$.

□

Lema 2 *Para cada $\epsilon>0$ e $m>0$ existe δ positivo tal que se $|x_0|<R$, $s_0\leq m$ então* $\mathrm{P}($*para alguns* $t, |x_t|<\epsilon)\geq\delta$.

Prova. Para que $|x_t|$ se aproxime da pequena vizinhança ϵ estudamos o comportamento do produtório

$$|x_t| = |x_0|\cdot\frac{|x_1|}{|x_0|}\cdots\frac{|x_t|}{|x_{t-1}|}.$$

Da equação (2.1) temos

$$\frac{x_t}{x_{t-1}} = 1-\gamma(s_{t-1})\frac{\varphi(x_{t-1})}{x_{t-1}} - \gamma(s_{t-1})\frac{\xi_t}{x_{t-1}}.$$

Suponhamos que $|x_{t-1}|\leq R$ e $|\xi_t|<\zeta_t$ com constantes $\{\zeta_t\}$ a serem especificadas. Então, usando que $\gamma(s_{t-1})\leq\gamma(0)$, $|\varphi(x_t)/x_t|\leq M$, obtemos

$$\frac{x_t}{x_{t-1}} \geq 1-\gamma(0)M-\gamma(0)\frac{\zeta_t}{\epsilon}. \tag{2.7}$$

Da Condição A3.2 segue que $1-\gamma(0)M>-1$. Por outro lado, denotando $g=\inf_{x\in[\epsilon,R]}\frac{\varphi(x)}{x}>0$ e usando que $s_{t-1}\leq m+(t-1)\mathrm{u}_+$, temos para $\epsilon\leq|x_{t-1}|\leq R$

$$\frac{x_t}{x_{t-1}} \leq 1-g\,\gamma(m+(t-1)\mathrm{u}_+)+\gamma(0)\frac{\zeta_t}{\epsilon}. \tag{2.8}$$

Denotamos por $G_t=\max\{|1-\gamma(0)M|, 1-g\,\gamma(m+(t-1)\mathrm{u}_+)\}$; obviamente que $G_t<1$. A divergência da série $\sum_t\gamma(m+t\mathrm{u}_+)$ implica que o produtório $\prod_{i=1}^{t-1}G_i$ tende para zero. De (2.7) e (2.8) segue que se $\epsilon\leq|x_{t-1}|\leq R$, $|x_t/x_{t-1}|\leq G_t+\gamma(0)\zeta_t/\epsilon$. Como $G_t\leq\sqrt{G_t}<1$, podemos escolher ζ_t tal que $G_t+\gamma(0)\zeta_t/\epsilon\leq\sqrt{G_t}$.

Assim,

$$|x_t/x_{t-1}|\leq\sqrt{G_t} \tag{2.9}$$

sempre que $\epsilon \leq |x_{t-1}| \leq R$ e $|\xi_t| < \zeta_t$. Escolhemos n tal que $R\prod_{t=1}^{n-1}\sqrt{G_t} < \epsilon$ e suponhamos que $|x_0| < R$, $s_0 \leq m$, e $|\xi_t| \leq \zeta_t$ enquanto $1 \leq t \leq n-1$. Então, para algum $t \in \{1, \ldots, n\}$, $|x_t| < \epsilon$.

Assim, pelo menos para o evento $\{\omega : |\xi_1| < \zeta_1, \ldots, |\xi_{n-1}| < \zeta_{n-1}\}$ fica assegurado o resultado de que para algum t, $|x_t| < \epsilon$. Então

$$\mathrm{P}(|\xi_1| < \zeta_1, \ldots, |\xi_{n-1}| < \zeta_{n-1}) = \mathrm{P}(|\xi_1| < \zeta_1) \cdots \mathrm{P}(|\xi_{n-1}| < \zeta_{n-1}) := \delta > 0.$$

□

Dos Lemas 1 e 2 vemos que para cada $\epsilon > 0$ existe $\delta > 0$ tal que para condições iniciais arbitrárias x_0, s_0, s_1

$$\mathrm{P}(\text{para alguns } t, |x_t| < \epsilon) > \delta.$$

Então podemos escolher um número inteiro positivo $n = n(x_0, s_0, s_1)$ tal que

$$\mathrm{P}(\text{para alguns } t \leq n, |x_t| < \epsilon) > \delta/2\,.$$

Denotamos $\bar{p} = \sup \mathrm{P}(\text{para cada } t, |x_t| \geq \epsilon)$, sendo o supremo tomado sobre todas as condições iniciais x_0, s_0, s_1. Fixamos x_0, s_0, s_1; então

$$\begin{aligned}
&\mathrm{P}(\text{para cada } t, |x_t| \geq \epsilon) = \\
&\quad = \mathrm{P}(\text{para cada } t > n, |x_t| \geq \epsilon \,\Big|\, \text{para cada } t \leq n, |x_t| \geq \epsilon) \cdot \mathrm{P}(\text{para cada } t \leq n, |x_t| \geq \epsilon) \leq \\
&\quad \leq \bar{p}\,(1 - \delta/2). \qquad (2.10)
\end{aligned}$$

O supremo do L.E. de (2.10) sobre todos os triplos (x_0, s_0, s_1) é $\bar{p}$. Assim obtemos a desigualdade $\bar{p} \leq \bar{p}\,(1 - \delta/2)$ de onde $\bar{p} = 0$. Obtemos pois o seguinte Lema:

Lema 3 *Para cada $\epsilon > 0$, quase-certamente existe t tal que $|x_t| < \epsilon$.*

Fixamos arbitrariamente ϵ e η positivos.

Lema 4 *Existe $\epsilon_1 > 0$ e $\delta > 0$ tal que se $|x_0| < \epsilon_1$ então*

$$\mathrm{P}(\textit{para alguns } t, |x_t| < \epsilon \textit{ e } s_t \geq \eta) \geq \delta.$$

Prova. Consideremos o evento

$$A = \{y_I \leq (-1)^t y_t \leq y_{II},\, t = 1, 2, \ldots, n-1\} \tag{2.11}$$

onde y_I, y_{II} e n serão especificados. Notemos que $y_I \leq (-1)^t y_t \leq y_{II}$ significa que $\xi_t \in I_t$ onde

$$I_t = \begin{cases} [y_I - \varphi(x_t), y_{II} - \varphi(x_t)] & \text{para } t \text{ ímpar} \\ [-y_{II} - \varphi(x_t), -y_I - \varphi(x_t)] & \text{para } t \text{ par .} \end{cases}$$

Vamos requerer que

$$y_{II} < \Omega/2 \tag{2.12}$$

e

$$|x_t| < \epsilon_2, \quad t = 1, 2, \ldots, n \tag{2.13}$$

onde ϵ_2 é escolhido por forma a que

$$\epsilon_2 < \epsilon \text{ and } \sup_{|x|<\epsilon_2} |\varphi(x)| < \Omega/2\,. \tag{2.14}$$

Então I_t pertence a $[-\Omega, \Omega]$, e de acordo com o Comentário 7, $\mathrm{P}(I_t) \geq p(y_{II} - y_I) > 0$. Assim $P(A) \geq \delta$ onde

$$\delta = [p(y_{II} - y_I)]^n. \tag{2.15}$$

Mais, temos $\mathrm{u}(-y_t y_{t-1}) \geq \mathrm{u}(y_I^2)$, $t = 1, 2, \ldots, n-1$, e então

$$s_t \geq t\,\mathrm{u}(y_I^2), \quad t = 1, 2, \ldots, n. \tag{2.16}$$

Mas além disso, para $t = 1, \ldots, n$

$$x_t = x_0 - \sum_{i=1}^{t} \gamma_{i-1} y_i, \text{ onde } \gamma_i = \gamma(s_i)\,. \tag{2.17}$$

A soma no L.D. de (2.17) pode ser estimada por

$$\begin{aligned} & y_I \sum_{i=1}^{t} (-1)^i \gamma_{i-1} - (y_{II} - y_I) \sum_{\substack{i=1 \\ (\text{par})}}^{t} \gamma_{i-1} \leq \\ \leq\ & \sum_{i=1}^{t} \gamma_{i-1} y_i \leq \\ \leq\ & y_{II} \sum_{i=1}^{t} (-1)^i \gamma_{i-1} + (y_{II} - y_I) \sum_{\substack{i=1 \\ (\text{ímpar})}}^{t} \gamma_{i-1}\,. \end{aligned} \tag{2.18}$$

O objectivo é obter as estimativas

$$\left| y_{II} \sum_{i=1}^{t} (-1)^i \gamma_{i-1} \right| < \epsilon_2/3 \tag{2.19}$$

e

$$\left| (y_{II} - y_I) \sum_{i=1}^{t} \gamma_{i-1} \right| < \epsilon_2/3, \tag{2.20}$$

então de (2.18) concluímos

$$\left| \sum_{i=1}^{t} \gamma_{i-1} y_i \right| < 2\epsilon_2/3$$

e colocando

$$\epsilon_1 = \epsilon_2/3, \tag{2.21}$$

de (2.17) garantimos que

$$|x_t| < \epsilon_2, \quad t = 1, \ldots, n. \tag{2.22}$$

Seja

$$y_{II} \leq \frac{\epsilon_2}{3\gamma(0)}, \tag{2.23}$$

então (2.19) fica garantida.

Agora, requeremos

$$y_I \geq y_{II}/2 \tag{2.24}$$

e escolhemos n tal que

$$n \geq \frac{\eta}{\mathrm{u}(y_{II}^2/4)}, \tag{2.25}$$

e de (2.16) concluímos que

$$s_n \geq \eta.$$

Usando (2.16) novamente, temos

$$\begin{aligned} \sum_{i=1}^{n} \gamma_{i-1} &= \gamma(s_0) + \gamma(s_1) + \sum_{i=2}^{n-1} \gamma(s_i) \leq \\ &\leq \quad 2\gamma(0) + \sum_{i=2}^{n-1} \gamma(i\,\mathrm{u}(y_{II}^2/4)) =: \Gamma(n) \end{aligned}$$

em que $\Gamma(n)$ é uma função crescente pela Condição A2.2. Assim, para garantir (2.20), devemos impor a condição

$$y_{II} - y_I \leq \frac{\epsilon_2}{3} \cdot \frac{1}{\Gamma(n)} \tag{2.26}$$

Agora, definimos y_{II} de acordo com (2.23) e (2.12); depois n, de acordo com (2.25); e finalmente, y_I, de acordo com (2.24) e (2.26). O valor δ é definido por (2.15). Então, provámos que $\mathrm{P}(|x_n| < \epsilon \text{ e } s_n \geq \eta) > \delta$.

□

Dos Lemas 3 e 4 segue para para cada $\epsilon > 0$ e $\eta > 0$ a probabilidade de que para alguns t, $|x_t| < \epsilon$ e $s_t \geq \eta$ seja maior que $\delta > 0$, depende somente de ϵ e η. Repetindo o argumento do Lema 3 temos

Lema 5 *Para cada $\epsilon > 0$ e $\eta > 0$, quase-certamente existe t tal que $|x_t| < \epsilon$ e $s_t \geq \eta$.*

Vamos definir o tempo-de-paragem $\tau(\epsilon) = \inf\{t : |x_t| \geq \epsilon\}$.

Lema 6 *Para cada $0 < \theta < \mathrm{E}_0$ existe uma constante $\epsilon_0 > 0$ e uma sequência π_n tal que $\lim_{n\to\infty} \pi_n = 0$ e*

$$\mathrm{P}(s_t > s_0 + t\theta - n \text{ para cada } t < \tau(\epsilon_0)) > 1 - \pi_n.$$

Prova. Vamos mostrar que

$$\mathrm{P}(\text{existe } t < \tau(\epsilon_0) \text{ tal que } s_t \leq s_0 + t\theta - n) \leq \pi_n \to 0\,.$$

De A4.2 segue que para algum ω_0 positivo existe $\mathrm{E}_{\omega_0} > \theta$ onde $\mathrm{E}_{\omega_0} = \mathrm{E}[u(X^{(\omega_0)})]$ e

$$X^{(\omega_0)} = \inf_{\substack{|\varphi_1|\leq\omega_0 \\ |\varphi_2|\leq\omega_0}} [-(\xi_1 + \varphi_1)(\xi_2 + \varphi_2)]. \tag{2.27}$$

Escolhemos ϵ_0 tal que

$$\sup_{|x|<\epsilon_0} |\varphi(x)| \leq \omega_0.$$

Definimos a sequência $\{\tilde{s}_t\}$ por

$$\tilde{s}_0 = s_0; \quad \tilde{s}_t = \tilde{s}_{t-1} + \mathrm{u}(X_t^{(\omega_0)}) \tag{2.28}$$

onde

$$X_t^{(\omega_0)} = \inf_{\substack{|\varphi_{t-1}|\leq\omega_0 \\ |\varphi_{t-2}|\leq\omega_0}} [-(\xi_t + \varphi_{t-1})(\xi_{t-1} + \varphi_{t-2})]. \tag{2.29}$$

Comparando (2.28) e (2.29) com (2.2), para $t < \tau(\epsilon_0)$, obtemos

$$\tilde{s}_t \leq s_t. \tag{2.30}$$

De (2.28) segue que

$$\tilde{s}_t - s_0 = t\mathrm{E}_{\omega_0} + \mathbb{I}_t^{\mathrm{par}} + \mathbb{I}_t^{\mathrm{ímpar}} \tag{2.31}$$

sendo

$$\mathbb{I}_t^{\mathrm{par}} = \sum_{\substack{i=1 \\ (i\ \mathrm{par})}}^{t} [\mathrm{u}(X_t^{(\omega_0)}) - \mathrm{E}_{\omega_0}], \quad \mathbb{I}_t^{\mathrm{ímpar}} = \sum_{\substack{i=1 \\ (i\ \mathrm{ímpar})}}^{t} [\mathrm{u}(X_t^{(\omega_0)}) - \mathrm{E}_{\omega_0}]$$

em que $\mathbb{I}_t^{\mathrm{par}}$ e $\mathbb{I}_t^{\mathrm{ímpar}}$ são somas de variáveis limitadas, independentes e identicamente distribuídas com média zero e variância linear com t.

Comentário 9 *Apesar de assimptoticamente normais, $\mathbb{I}_t^{par}$ e $\mathbb{I}_t^{ímpar}$ são dependentes entre si pelo que usamos o seguinte argumento para estimar a probabilidade da sua soma: $X + Y < a$ implica que $X < a/2$ ou $Y < a/2$ onde X e Y são variáveis aleatórias reais e a uma constante real. Assim*

$$\mathrm{P}(X + Y < a) \leq \mathrm{P}(X < a/2) + \mathrm{P}(Y < a/2) \simeq 2\mathrm{P}(X < a/2).$$

Então, por $\mathrm{Var}\, \mathbb{I}_t^{\mathrm{par}} = t \cdot V_{\mathbb{I}_1}$, temos

$$\mathrm{P}(\mathbb{I}_t^{\mathrm{par}} + \mathbb{I}_t^{\mathrm{ímpar}} < 2a) \lesssim 2\mathrm{P}(\mathbb{I}_t^{\mathrm{par}} < a) \leq 2\Phi(\frac{a}{\sqrt{t}\sqrt{V}_{\mathbb{I}_1}}). \tag{2.32}$$

Do evento $s_t \leq s_0 + t\theta - n$, sabendo que $\tilde{s}_t \leq s_t$ para $t < \tau(\epsilon_0)$, segue

$$\begin{aligned} \tilde{s}_t &\leq s_0 + t\theta - n \Leftrightarrow \\ s_0 + t\mathrm{E}_{\omega_0} + \mathbb{I}_t^{\mathrm{par}} + \mathbb{I}_t^{\mathrm{ímpar}} &\leq s_0 + t\theta - n \Leftrightarrow \\ \mathbb{I}_t^{\mathrm{par}} + \mathbb{I}_t^{\mathrm{ímpar}} &\leq -t(\mathrm{E}_{\omega_0} - \theta) - n \end{aligned} \tag{2.33}$$

Comentário 10 *No que se segue usamos a seguinte expressão, onde $\{X_i, i = 1, \ldots, \infty\}$ é uma sequência de variáveis aleatórias,*

$$\mathrm{P}(\textit{existe } t < \tau \textit{ tal que } X_t < a) \leq \sum_{i=1}^{\tau} \mathrm{P}(X_i < a) \leq \sum_{i=1}^{\infty} \mathrm{P}(X_i < a) \tag{2.34}$$

Por (2.32), (2.33) e (2.34) segue

$$\begin{aligned} \mathrm{P}(\text{existe } t < \tau(\epsilon_0) \text{ tal que } s_t \leq s_0 + t\theta - n) &\leq \\ \mathrm{P}(\text{existe } t < \tau(\epsilon_0) \text{ tal que } \mathbb{I}_t^{\mathrm{par}} + \mathbb{I}_t^{\mathrm{ímpar}} \leq -t(\mathrm{E}_{\omega_0} - \theta) - n) &\leq \\ \sum_{i=1}^{\infty} \mathrm{P}(\mathbb{I}_i^{\mathrm{par}} + \mathbb{I}_i^{\mathrm{ímpar}} \leq -i(\mathrm{E}_{\omega_0} - \theta) - n) &\lesssim \\ 2\sum_{i=1}^{\infty} \mathrm{P}(\frac{\mathbb{I}_i^{\mathrm{par}}}{\sqrt{iV_I}} \leq -\sqrt{i}\frac{\mathrm{E}_{\omega_0} - \theta}{\sqrt{V}_I} - \frac{n}{\sqrt{iV_I}}) &\leq \\ 2\sum_{i=1}^{\infty} \Phi(-\sqrt{i}K_1 - \frac{n}{\sqrt{i}}K_2) &:= \pi_n \end{aligned}$$

para certas constantes $K_1 > 0$ e $K_2 > 0$. A série da última desigualdade é convergente pelo que $\pi_n \to 0$ e então

$$\begin{aligned}\pi_n := \mathrm{P}(&\text{existe } t \text{ tal que } \mathbb{I}_t^{\text{par}} + \mathbb{I}_t^{\text{ímpar}} \leq \\ &\leq -n - t(\mathrm{E}_{\omega_0} - \theta)) \to 0 \text{ quando } n \to \infty.\end{aligned}$$

□

Fixamos θ e ϵ_0 como no Lema 6, e escolhemos arbitrariamente $\epsilon < \epsilon_0$ positivo e também n. Definimos o tempo-de-paragem $\nu = \nu(n, \epsilon) = \inf\{t : |x_t| \geq \epsilon \text{ ou } s_t \leq s_0 - n + t\theta\}$, e escolhemos $\epsilon_1 > 0$ tal que $\sup_{|x|<\epsilon_1} f(x) < \frac{1}{2} \inf_{|x|>\epsilon} f(x)$. (Relembramos que f é a primitiva de φ satisfazendo $f(0) = 0$.)

Lema 7 *Seja $|x_0| < \epsilon_1$, então $\mathrm{P}(\nu < \infty) \leq K \int_{s_0-n-1}^{\infty} \gamma^2(s)ds + \pi_n$, sendo K uma constante que depende apenas de ϵ.*

Prova. Tomando as notações f_t, f'_t, γ_t do Lema 1 e usando (2.5), temos

$$f_t - f_{t-1} \leq -\gamma_{t-1} f'_{t-1} \xi_t + M\gamma_{t-1}^2 ({f'}_{t-1}^2 + \xi_t^2).$$

Isto implica que

$$f_t - f_0 \leq \mathbb{I}'_t + \mathbb{I}''_t$$

onde

$$\mathbb{I}'_t = \left| \sum_{i=1}^{t} \gamma_{i-1} f'_{i-1} \xi_i \right|, \qquad \mathbb{I}''_t = M \sum_{i=1}^{t} \gamma_{i-1}^2 ({f'_{i-1}}^2 + \xi_i^2).$$

Denotando $P' = \mathrm{P}(\, \mathbb{I}'_\nu \cdot \mathbb{I}(\nu < \infty) \geq \delta/2)$, $P'' = \mathrm{P}(\, \mathbb{I}''_\nu \cdot \mathbb{I}(\nu < \infty) \geq \delta/2)$ com $\delta = \frac{1}{2} \inf_{|x|\geq\epsilon} f(x)$ e usando os Lemas 5 e 6 obtemos

$$\begin{aligned}\mathrm{P}(\nu < \infty) &\leq \mathrm{P}(s_t \leq s_0 - n + t\theta) + \mathrm{P}(|x_t| \geq \epsilon) \\ &\leq \pi_n + \mathrm{P}(f_t f_0 > \delta) \\ &\leq \pi_n + \mathrm{P}(I' > \delta/2) + \mathrm{P}(I'' > \delta/2) \\ &\leq \pi_n + P' + P''\end{aligned} \tag{2.35}$$

De acordo com a desigualdade de Markov (por exemplo, [48, p. 59]),

$$\begin{aligned}P' &\leq \frac{4}{\delta^2} \mathrm{E}[\, {\mathbb{I}'_\nu}^2 \cdot \mathbb{I}^2(\nu < \infty)] = \\ &= \frac{4}{\delta^2} \sum_{i,j=1}^{\infty} \mathrm{E}[\gamma_{i-1} f'_{i-1} \xi_i I(i-1 < \nu < \infty) \cdot \gamma_{j-1} f'_{j-1} \xi_j I(j-1 < \nu < \infty] \leq \\ &\leq \frac{4}{\delta^2} \sum_{i,j=1}^{\infty} \mathrm{E}[\gamma_{i-1} f'_{i-1} \xi_i I(i-1 < \nu) \cdot \gamma_{j-1} f'_{j-1} \xi_j I(j-1 < \nu].\end{aligned} \tag{2.36}$$

Relembrando que o conjunto de variáveis $\{\gamma_{i-1}, f'_{i-1},\ \mathbb{I}(i-1<\nu)\}$ e ξ_i é mutuamente independente, obtemos facilmente que os termos no L.D. de (2.36) com $i \neq j$ são iguais a zero, e então

$$P' \leq \frac{4}{\delta^2}\sum_{i=1}^{\infty} \mathrm{E}[\gamma_{i-1}^2\, {f'_{i-1}}^2\ \mathbb{I}(i<\nu)\xi_i^2] \leq K'\,\mathrm{E}\left[\sum_{i=1}^{\nu-1}\gamma_{i-1}^2\right] \tag{2.37}$$

onde $K' = \frac{4S^2}{\delta^2}\sup_{|x|<\epsilon} f'^2(x)$.

Analogamente,

$$P'' \leq \frac{2M}{\delta}\mathrm{E}\left[\sum_{i=1}^{\infty}\gamma_{i-1}^2({f'_{i-1}}^2+\xi_i^2)\cdot\ \mathbb{I}(i-1<\nu)\right] \leq K''\mathrm{E}\left[\sum_{i=1}^{\nu-1}\gamma_{i-1}^2\right] \tag{2.38}$$

com $K'' = (2M/\delta)(\sup_{|x|<\epsilon} f'^2(x) + S^2)$.

Para $t<\nu$, $s_t > s_0 + t\theta - n$, temos que $\gamma_t < \gamma(s_0 - n + t\theta)$, e

$$\mathrm{E}\left[\sum_{i=1}^{\nu-1}\gamma_{i-1}^2\right] < \sum_{i=1}^{\infty}\gamma^2(s_0-n+i\theta) \leq \frac{1}{\theta}\int_{s_0-n-1}^{\infty}\gamma^2(s)ds. \tag{2.39}$$

Tomando $K = \theta^{-1}(K'+K'')$, de (2.35), (2.37), (2.38) e (2.39) obtemos o Lema 7.

□

Agora, fixamos $\epsilon < \epsilon_0$ positivo e escolhemos n e η tal que $1 - \pi_n - K\int_{\eta-n-1}^{\infty}\gamma^2(s)ds =: \delta$ seja positivo. Fixamos também $\epsilon_1 = \epsilon_1(\epsilon)$ como definido acima. De acordo com os Lemas 5 e 7, *quase-certamente* existe t_0 tal que $|x_{t_0}| < \epsilon_1$, $s_{t_0} \geq \eta$, e a probabilidade para todo o $t \geq t_0$, $|x_t| < \epsilon$ excede δ.

Definimos a sequência de tempos-de-paragem $\tau_1 = 1$,

$$\tau_{i+1} = \inf\{\tau > \tau_i : |x_\tau| \geq \epsilon,\ \text{e para algum } \tau_i \leq t < \tau,\ |x_t| < \epsilon_1 \text{ e } s_t > \eta\}, \quad i = 1, 2, \ldots .$$

Temos

$$\mathrm{P}(\tau_{i+1} = \infty \,|\, \tau_i < \infty) \geq \delta,$$

de onde

$$\mathrm{P}(\tau_{i+1} < \infty) = \mathrm{P}(\tau_{i+1} < \infty \,|\, \tau_i < \infty)\,\mathrm{P}(\tau_i < \infty) \leq (1-\delta)\,\mathrm{P}(\tau_i < \infty).$$

Assim, $\mathrm{P}(\tau_i < \infty) \to 0$ as $i \to \infty$; isto implica que *quase-certamente* $i_0 = \sup\{i : \tau_i < \infty\}$ é finito.

Como, de acordo com o Lema 5, *quase-certamente* existe $t_0 \geq \tau_{i_0}$ tal que $|x_{t_0}| < \epsilon_1$ e $s_{t_0} > \eta$, concluímos que $|x_t| < \epsilon$ quando $t > t_0$. O Teorema 1 está demonstrado. □

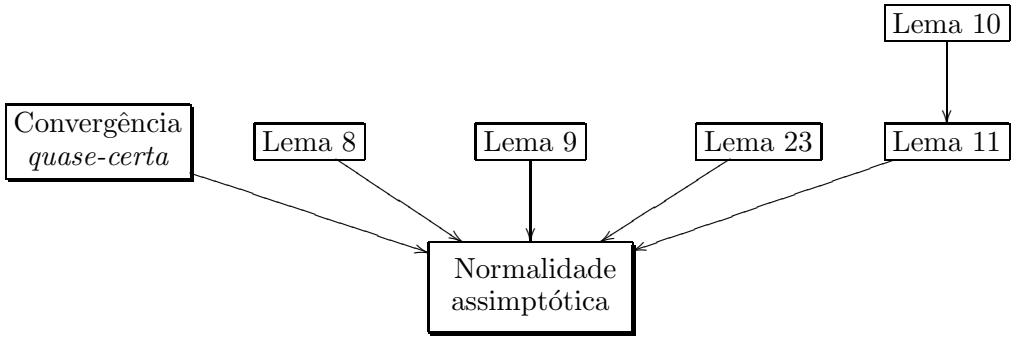


Figura 2.3: Lemas para a demonstração da normalidade assimptótica.

2.3 Demonstração da Normalidade Assimptótica

A demonstração da normalidade assimptótica segue o trabalho de Delyon e Juditsky [14] e será referido quando estivermos a usar partes da demonstração original. A organização da demonstração está ilustrada na Figura 2.3 onde os passos principais da demonstração são:

- O Lema 8 descreve o comportamento assimptótico da variável s_t.

- Consideremos z_t o processo resultante de aplicar o algoritmo de Robbins-Monroe a uma função $\varphi(x) = \alpha x$ e para o qual, os resultados de normalidade assimptótica de $\sqrt{t}z_t$ são conhecidos. O Lema 11 mostra que $\sqrt{t}(x_t - z_t)$ converge em probabilidade em que x_t é o processo resultante da aplicação do algoritmo de Kesten Generalizado a uma função φ cuja derivada em x^* é $\varphi'(x^*) = \alpha$.

- A conclusão da demonstração conforme o trabalho original.

Consideramos, sem perda de generalidade, que $x^* = 0$.

Recordamos a definição de E_0 na Condição A4.2. Recordamos a definição de E_0 na Condição A4.2.

Lema 8 *Sejam s_0 e s_1 variáveis aleatórias que são condição inicial do processo $\{s_t\}$, definido em (2.2). Então*

$$\gamma(s_t) = 1/s_t = \frac{1}{\mathrm{E}_0 t}(1 + \mathrm{o}_t), \ \textit{quase-certamente} \tag{2.40}$$

onde o_t é uma v.a. que depende de $\{\xi_i, i \leq t\}$ de das condições iniciais e que verifica $\lim_{t\to\infty} \mathrm{o}_t = 0$ quase-certamente.

Prova. A Condição A4.5 permite a decomposição

$$\begin{aligned} \mathrm{u}(-y_{i-1}y_i) &= \mathrm{u}(-(\varphi_{i-2}+\xi_{i-1})(\varphi_{i-1}+\xi_i)) = \\ &= \mathrm{u}(-(\varphi_{i-2}+\xi_{i-1})(\varphi_{i-1}+\xi_i)) = \\ &= \mathrm{u}(-\varphi_{i-2}\varphi_{i-1}-\varphi_{i-2}\xi_i-\varphi_{i-1}\xi_{i-1}-\xi_{i-1}\xi_i) = \\ &= \mathrm{u}(-\xi_{i-1}\xi_i)+\mathrm{u}'(\theta_i)\times(-\varphi_{i-2}\varphi_{i-1}-\varphi_{i-2}\xi_i-\varphi_{i-1}\xi_{i-1}) \end{aligned} \tag{2.41}$$

onde θ_i é um ponto entre $-y_{i-1}y_i$ e $-\xi_{i-1}\xi_i$. Temos também que a função u' é limitada e $\varphi(x_i)\to 0$ de onde, pelo corolário do Lema de Kronecker (por exemplo, [48]),

$$\sum_{i=1}^{t}\mathrm{u}'(\theta_i)\varphi_{i-2}\varphi_{i-1} = o(t), \tag{2.42}$$

$$\sum_{i=1}^{t}\mathrm{u}'(\theta_i)\varphi_{i-2}\xi_i = o(t), \tag{2.43}$$

$$\sum_{i=1}^{t}\mathrm{u}'(\theta_i)\varphi_{i-1}\xi_{i-1} = o(t)\,. \tag{2.44}$$

Assim temos

$$\begin{aligned} s_t &= s_0+s_1+\sum_{i=1}^{t}(\mathrm{u}(-y_{i-1}y_i)-\mathrm{u}(-\xi_{i-1}\xi_i))+ \\ &\quad +\sum_{\text{pares}}^{t}\mathrm{u}(-\xi_{i-1}\xi_i)+\sum_{\text{ímpares}}^{t}\mathrm{u}(-\xi_{i-1}\xi_i) \\ &= s_0+s_1+\Delta U_t+P_t+I_t. \end{aligned}$$

Por (2.42), (2.43) e (2.44)

$$\Delta U_t=\sum_{i=1}^{t}(\mathrm{u}(-y_{i-1}y_i)-\mathrm{u}(-\xi_{i-1}\xi_i))=o(t)\ \textit{quase-certamente}.$$

Cada uma das somas P_t e I_t é composta por parcelas independentes de média E_0 e variância finita. Pela lei do logaritmo iterado

$$P_t+I_t=\mathrm{E}_0t+\mathrm{O}(\sqrt{t\log\log t})\,.$$

Sendo $\lim_{t\to\infty}s_0/t=0$ *quase-certamente*, idem para s_1, temos

$$s_t=s_0+s_1+\mathrm{E}_0t+t\mathrm{o}_t+\mathrm{O}(\sqrt{t\log\log t})=(\mathrm{E}_0+\mathrm{o}_t)t,$$

quase-certamente. Então

$$\begin{aligned} s_t &= (\mathrm{E}_0+\mathrm{o}_t)t=\mathrm{E}_0t\left(\frac{1}{1-\frac{\mathrm{o}_t}{\mathrm{E}_0+\mathrm{o}_t}}\right)= \\ &= \mathrm{E}_0t\left(\frac{1}{1+\mathrm{o}_t}\right). \end{aligned}$$

□

Lema 9 (Delyon e Juditsky [14]) *Seja (ν_t) uma sequência aleatória de números reais tais que $\nu_t \to 0$ quase-certamente quando $t \to \infty$. Então existe uma sequência determinística (a_t) tal que*

$$a_t \to 0 \quad e \quad \nu_t/a_t \to 0 \quad \textit{quase-certamente.} \tag{2.45}$$

A próxima demonstração segue o trabalho análogo de Delyon e Juditsky [14] contendo uma adaptação ao novo algoritmo.

Conclusão da demonstração do Teorema 2. Fazemos $x^* = 0$. A convergência *quase-certa* de $x_t \to 0$, o Lema 8 e o Lema 9 conduzem a que exista uma sequência (a_t) de números não aleatórios positivos tal que

$$a_t \to 0 \quad \text{e} \quad |\mathrm{o}_t|/a_t \to 0, \quad |x_t|/a_t \to 0 \quad \textit{quase-certamente.} \tag{2.46}$$

Comentário 11 *A explicação para o acima dado é que se fizermos $\theta_t := |o_t| + |x_t|$ então $\theta_t \to 0$ quase-certamente. Então existe $a_t \to 0$, deterministicamente, tal que $\theta_t/a_t \to 0$ quase-certamente. Daqui segue que $|o_t|/a_t \to 0$ e $|x_t|/b_t \to 0$ quase-certamente.*

Definimos os tempos-de-paragem

$$\tau_R = \inf\{t : |\mathrm{o}_t| \geq R|a_t|\}, \quad \sigma_R = \inf\{t : |x_t| \geq R|a_t|\} \tag{2.47}$$

para $R > 0$ e

$$\nu = \min(\tau_R, \sigma_R)\,. \tag{2.48}$$

Do Lema 9 e de (2.46) concluímos que para qualquer $\epsilon > 0$ podemos escolher $R < \infty$ tal que

$$\mathrm{P}(\nu = \infty) \geq 1 - \epsilon. \tag{2.49}$$

Comentário 12 *Desta maneira, com uma probabilidade tão grande quanto se deseje, temos um majorante determinístico comum a $|o_t|$ e $|x_t|$ e que será usado na demonstração.*

Consideramos o processo análogo ao algoritmo em (2.1) mas com passo determinístico $\gamma_t = 1/(\mathrm{E}_0 t)$ e aplicado à função $\varphi(x) = \alpha x$ (α é a derivada de φ em x^*),

$$z_t = z_{t-1} - \frac{1}{\mathrm{E}_0 t}(\alpha z_{t-1} + \xi_t), \quad z_0 = x_0. \tag{2.50}$$

As propriedades assimptóticas deste processo são conhecidas (ver, por exemplo, Nevel'son e Has'minskii [31]). Assim

$$\begin{gathered} z_t t^{1/2-\epsilon} \to 0, \textit{quase-certamente}, \text{ para todo } \epsilon > 0, \\ \mathrm{E} z_t^2 \leq K/t, \quad K > 0 \\ \sqrt{t} z_t \xrightarrow{\mathrm{d}} N\left(0, \frac{S^2}{\mathrm{E}_0(2\alpha - \mathrm{E}_0)}\right). \end{gathered} \tag{2.51}$$

Baseado no Lema 23, pág. 62, o Lema 11 demonstrará que, assimptoticamente, $\sqrt{t}x_t$ e $\sqrt{t}z_t$ têm a mesma distribuição limite, descrita em (2.51). □

Lema 10 *Consideremos a seguinte expressão, em que $b > 0$, a_0 é um número real,*

$$0 \leq a_{t+1} \leq (1 - \frac{b}{t})a_t + \mathrm{o}(t^{-1}), \quad t = 1, 2, \ldots . \tag{2.52}$$

Então $a_t \to 0$.

Prova. Considaremos a sequência recursiva, onde ϵ é um número real positivo,

$$0 \leq A_{t+1} \leq (1 - \frac{b}{t})A_t + \epsilon/t, \quad t = t_0, t_0 + 1, \ldots .$$

e transformando

$$0 \leq A_{t+1} \leq A_t - \frac{bA_t - \epsilon}{t}, \quad t = t_0, t_0 + 1, \ldots .$$

ou

$$0 \leq bA_{t+1} - \epsilon \leq bA_t - \epsilon - b\frac{bA_t - \epsilon}{t}, \quad t = t_0, t_0 + 1, \ldots .$$

Escrevemos $B_t = bA_t - \epsilon$ e temos

$$B_{t+1} = B_t(1 - b/t)$$

e então $B_t \to 0$, pelo que $A_t \to \epsilon/b$.

A sequência enunciada no lema é

$$0 \leq a_{t+1} \leq (1 - \frac{b}{t})a_t + \mathrm{o}(1)/t, \quad t = 1, 2, \ldots .$$

para a qual escolhemos $\epsilon > 0$ tal que $o(1) < \epsilon$ se $t \geq t_0$ para algum t_0. Definimos

$$A_{t+1} = (1 - \frac{b}{t})A_t + \epsilon/t, \quad t = t_0, t_0 + 1, \ldots .$$

e $A_{t_0} = a_{t_0}$. Usando indução supomos $A_t - a_t \geq 0$ para $t \geq t_0$. Para $t + 1$

$$A_{t+1} - a_{t+1} = (1 - \frac{b}{t})(A_t - a_t) + (\epsilon - o(1))/t$$

verificando que $A_{t+1} - a_{t+1} \geq 0$ usando a hipótese. Assim $0 \leq a_t \leq A_t$.

Sendo $A_t \to \epsilon/b$ e dado que podemos escolher ϵ tão pequeno quando o desejado, concluímos que $A_t \to 0$ e daqui segue que $a_t \to 0$.

□

Lema 11 *Seja $\Delta_t = x_t - z_t$. Então $\sqrt{t}\Delta_t \xrightarrow{pr} 0$. ($\xrightarrow{pr}$ denota convergência em probabilidade.)*

Prova. De acordo com a Condição A4.4, a função φ admite decomposição de Taylor em que

$$\varphi(x_t) = \varphi'(0)x_t + \frac{1}{2}\varphi''(\theta_t x_t)x_t^2 \tag{2.53}$$

algum θ_t em $(0,1)$ e em que $|\varphi''(\theta_t x_t)| \leq D$ (ver A4.4). Denotamos $e_t := \frac{1}{2}\varphi''(\theta_t x_t)$.

Da definição de x_{t+1}, (2.1)–(2.2), Lema 8 e (2.50), obtemos

$$\begin{array}{rcl} x_{t+1} & = & x_t - \dfrac{\alpha x_t}{\mathrm{E}_0 t}(1+\mathrm{o}_t) - \dfrac{e_t x_t^2}{\mathrm{E}_0 t}(1+\mathrm{o}_t) - \dfrac{\xi_{t+1}}{\mathrm{E}_0 t}(1+\mathrm{o}_t) \\ z_{t+1} & = & z_t - \dfrac{\alpha z_t}{\mathrm{E}_0 t} - \dfrac{\xi_{t+1}}{\mathrm{E}_0 t} \end{array} \tag{2.54}$$

de onde

$$\begin{array}{rcl} \Delta_{t+1} & = & \Delta_t(1 - \dfrac{\alpha}{\mathrm{E}_0 t}) - \dfrac{\alpha x_t}{\mathrm{E}_0 t}\mathrm{o}_t - \dfrac{e_t x_t^2}{\mathrm{E}_0 t}(1+\mathrm{o}_t) - \dfrac{\xi_{t+1}}{\mathrm{E}_0 t}\mathrm{o}_t \\ & =: & \Delta_t A_t - B_t - C_t - D_t\,. \end{array} \tag{2.55}$$

Usando $(a+b+c+d)^2 \leq a^2 + 3(b^2+c^2+d^2) + 2a(b+c+d)$ (obtém-se de desigualdades como $bc + cb \leq b^2 + c^2$)

$$\Delta_{t+1}^2 \leq \Delta_t^2 A_t^2 + 3(B_t^2 + C_t^2 + D_t^2) + 2\Delta_t A_t(-B_t - C_t - D_t)\,. \tag{2.56}$$

Considerando $t \leq \nu$ (definido em (2.48),

$$A_t^2 = (1 - \frac{\alpha}{\mathrm{E}_0 t})^2 \tag{2.57}$$

$$3B_t^2 = 3(\frac{\alpha^2 x_t}{\mathrm{E}_0 t}\mathrm{o}_t)^2 \leq 3\frac{\alpha R^2 a_t^2}{\mathrm{E}_0^2 t^2} R^2 a_t^2 = \frac{3\alpha^2 R^4}{\mathrm{E}_0}\frac{a_t^4}{t^2} \tag{2.58}$$

$$3C_t^2 = 3\left(\frac{e_t x_t^2}{\mathrm{E}_0 t}(1+\mathrm{o}_t)\right)^2 \leq 3\left(\frac{e_t^2 x_t^4}{\mathrm{E}_0^2 t^2}(2+2\mathrm{o}_t^2)\right) \leq \tag{2.59}$$

$$\leq \frac{6D^2R^4}{\mathrm{E}_0^2}\frac{a_t^4}{t^2} + \frac{6D^2R^6}{\mathrm{E}_0^2}\frac{a_t^6}{t^2} \tag{2.60}$$

$$3D_t^2 = 3\left(\frac{\xi_{t+1}\,\mathrm{o}_t}{\mathrm{E}_0 t}\right)^2 = 3\frac{\xi_{t+1}^2 \mathrm{o}_t^2}{\mathrm{E}_0^2 t^2} \leq \frac{3\xi_{t+1}^2}{\mathrm{E}_0^2}\frac{R^2 a_t^2}{t^2}\,. \tag{2.61}$$

Usando $-2ab \leq a^2 + b^2$

$$\begin{aligned} -2\Delta_t x_t &\leq \Delta_t^2 + x_t^2 \\ &= \Delta_t^2 + (\Delta_t + z_t)^2 \\ &\leq \Delta_t^2 + 2\Delta_t^2 + 2z_t^2 \\ &= 3\Delta_t^2 + 2z_t^2 \end{aligned}$$

e continuando com $t \leq \nu$

$$\begin{aligned} 2\Delta_t A_t(-B_t) &= 2\Delta_t(1 - \frac{\alpha}{\mathrm{E}_0 t})\frac{\alpha(-x_t)}{\mathrm{E}_0 t}\mathrm{o}_t = \\ &= (3\Delta_t^2 + 2z_t^2)(1 - \frac{\alpha}{\mathrm{E}_0 t})\frac{\alpha}{\mathrm{E}_0 t}\mathrm{o}_t = \\ &\leq (3\Delta_t^2 + 2z_t^2)\frac{\alpha}{\mathrm{E}_0 t}Ra_t \leq \\ &\leq \frac{3\alpha Ra_t}{\mathrm{E}_0 t}\Delta_t^2 + z_t^2\frac{2\alpha R}{\mathrm{E}_0 t}a_t \end{aligned} \tag{2.62}$$

e com o majorante $-e_t \leq D$,

$$\begin{aligned} 2\Delta_t A_t(-C_t) &= 2\Delta_t(1 - \frac{\alpha}{\mathrm{E}_0 t})\frac{-e_t x_t^2}{\mathrm{E}_0 t}(1 + \mathrm{o}_t) = \\ &= (3\Delta_t^2 + 2z_t^2)(1 - \frac{\alpha}{\mathrm{E}_0 t})\frac{-e_t x_t}{\mathrm{E}_0 t}(1 + \mathrm{o}_t) \leq \\ &\leq (3\Delta_t^2 + 2z_t^2)\frac{DRa_t}{\mathrm{E}_0 t}(1 + Ra_t) \leq \\ &\leq \frac{3DRa_t}{\mathrm{E}_0 t}(1 + Ra_t)\Delta_t^2 + z_t^2\frac{2DRa_t}{\mathrm{E}_0 t}(1 + Ra_t) \end{aligned} \tag{2.63}$$

$$2\Delta_t A_t(-D_t) = 2\Delta_t(1 - \frac{\alpha}{\mathrm{E}_0 t})\frac{-\xi_{t+1}}{\mathrm{E}_0 t}\mathrm{o}_t \tag{2.64}$$

De (2.56) e (2.62) os termos em Δ_t^2 são

$$\begin{aligned} &\left(1 - \frac{\alpha}{\mathrm{E}_0 t}\right)^2 + \frac{3\alpha Ra_t}{\mathrm{E}_0 t} + \frac{3DRa_t}{\mathrm{E}_0 t}(1 + Ra_t) \leq \\ &= 1 - \frac{2\alpha/\mathrm{E}_0 - 3\alpha Ra_t/\mathrm{E}_0 - (3DRa_t)(1 + Ra_t)/\mathrm{E}_0 - (\alpha^2/\mathrm{E}_0^2)/t}{t} \\ &\leq 1 - \frac{2\alpha/\mathrm{E}_0 - \mathrm{o}(1)}{t} \end{aligned} \tag{2.65}$$

onde o(1) é um infinitésimo.

De (2.58) a (2.61) temos os termos de $1/t^2$

$$\mathrm{o}(t^{-2}) = \frac{1}{t^2}\frac{3\alpha^2 R^4 a_t^4}{\mathrm{E}_0} + \frac{6D^2 R^4 a_t^4}{\mathrm{E}_t^2} + \frac{6D^2 R^6 a_t^6}{\mathrm{E}_0^2}. \tag{2.66}$$

De 2.62 e de 2.63 temos os termos em z_t^2

$$z_t^2 \mathrm{o}(1) = z_t^2 \frac{2\alpha R a_t}{\mathrm{E}_0} + \frac{2DRa_t}{\mathrm{E}_0}(1 + Ra_t). \tag{2.67}$$

De (2.64) temos o termo em ξ_{t+1}^2

$$\xi_{t+1}^2 \mathrm{o}(1) = \xi_{t+1}^2 \frac{3R^2 a_t^2}{\mathrm{E}_0^2}. \tag{2.68}$$

Com $\mathbb{I}(t+1 \leq \nu) \leq \mathbb{I}(t \leq \nu)$,

$$\begin{aligned} \Delta_{t+1}^2 \mathbb{I}(t+1 \leq \nu) \quad \leq \quad & \Delta_t^2 \mathbb{I}(t \leq \nu)(1 - \frac{2\alpha/\mathrm{E}_0 - \mathrm{o}_t^*}{t}) \\ & + \mathrm{o}(t^{-2}) \mathbb{I}(t \leq \nu) \\ & + \mathrm{o}(t^{-1}) z_t^2 \mathbb{I}(t \leq \nu) \\ & + \mathrm{o}(t^{-2}) \xi_{t+1}^2 \\ & + 2\Delta_t (1 - \frac{\alpha}{\mathrm{E}_0 t}) \frac{-\xi_{t+1}}{\mathrm{E}_0 t} \mathrm{o}_t \mathbb{I}(t \leq \nu) \end{aligned} \tag{2.69}$$

e porque $\mathrm{E}(z_t^2 \mathbb{I}(t \leq \nu)) \leq K_1/t$, $\mathrm{E}(\xi_{t+1} \mathbb{I}(t \leq \nu)) = 0$ obtemos

$$\mathrm{E}\Delta_{t+1}^2 \mathbb{I}(t+1 \leq \nu) \leq \mathrm{E}\Delta_t^2 \mathbb{I}(t \leq \nu)(1 - \frac{2\alpha/\mathrm{E}_0 - \mathrm{o}(1)}{t}) + \mathrm{o}(t^{-2}) \tag{2.70}$$

Vamos mostrar a convergência em média quadrática de $t\mathrm{E}\Delta_t^2 \to 0$ e pelo Teorema 7, pág. 62, vamos verificar que $\sqrt{t}(x_t - z_t) \xrightarrow{\mathrm{pr}} 0$. Definimos

$$W_t := (t+1)\mathrm{E}\Delta_{t+1}^2 \mathbb{I}(t+1 \leq \nu) \leq (t+1) \left(\mathrm{E}\Delta_t^2 \mathbb{I}(t \leq \nu)(1 - \frac{2\alpha/\mathrm{E}_0 - \mathrm{o}(1)}{t}) + \mathrm{o}(t^{-2}) \right) \tag{2.71}$$

Temos $(t+1)\mathrm{o}(t^{-2}) = \mathrm{o}(t^{-1})$ e para se poder usar o resultado do Lema 10 fazemos

$$\begin{aligned} (t+1)(1 - \frac{2\alpha/\mathrm{E}_0 - \mathrm{o}(1)}{t}) \quad &= \quad t \times (1 - \frac{H_t}{t}) \Leftrightarrow \\ H_t \quad &= \quad 2\alpha/\mathrm{E}_0 - 1 + \mathrm{o}(t^{-1}) \end{aligned}$$

de onde concluímos que $H_t > 0$ a partir de certo t observando a Condição A4.3. Como $\mathrm{E}W_1 < \infty$ porque $\mathrm{E}x_0 < \infty$ então, para certo $H > 0$, pelo Lema 10 obtemos $W_t \to 0$.

Daqui segue que

$$\sqrt{t}\Delta_t \mathbb{I}(t < \nu) \xrightarrow{\mathrm{pr}} 0. \tag{2.72}$$

A demonstração de (2.72) é válida para qualquer R e porque o evento $\mathbb{I}(t < \nu) = 1$ pode ocorrer com probabilidade arbitrariamente grande concluímos que $\sqrt{t}\Delta_t \xrightarrow{\mathrm{pr}} 0$.

□

2.4 Condições e Teoremas para o Caso Multidimensional

É demonstrada a convergência quase-certa e a normalidade assimptótica duma generalização do algoritmo de aproximação estocástica de Kesten para o caso multidimensional. A demonstração da convergência segue as linhas do caso unidimensional Plakhov e Cruz (2004) [35] e a demonstração da normalidade assimptótica segue o trabalho de Delyon e Judistky (1993) [14].

Consideramos o problema de encontrar o ponto de estacionaridade $x^* \in \mathbb{R}^n$ dum campo vectorial $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ de acordo com o algoritmo de Aproximação Estocástica

$$x_t = x_{t-1} - \gamma(s_{t-1})y_t, \quad t = 1, 2, \ldots \tag{2.73}$$

$$s_t = (s_{t-1} + \mathrm{u}(-y_t^T y_{t-1}))^+, \quad t = 2, 3, \ldots \tag{2.74}$$

onde

- $y_t = \varphi(x_{t-1}) + \xi_t$, $y_t \in \mathbb{R}^n$ é a t–ésima medida de φ perturbada do vector aleatório $\xi_t \in \mathbb{R}^n$;
- $a^+ := \max\{a, 0\}$;
- u é uma função sigmóide;
- O vector aleatório $x_0 \in \mathbb{R}^n$, e as variáveis aleatórias s_0 e s_1 são condições iniciais do algoritmo;
- $x_t \in \mathbb{R}^n$ é a t–ésima aproximação ao ponto de estacionaridade $x^* \in \mathbb{R}^n$ de φ.

Supomos que as condições seguintes se verificam.

Condições B1

1. $\{x_0, \xi_1, \xi_2, \ldots, \}$ são vectores aleatórios mutuamente independentes em que os vectores ξ_i são identicamente distribuídos com média zero $\mathrm{E}\xi_t = 0$ e covariâncias finitas $S_\xi := \mathrm{E}\,\xi_t\xi_t^T$.
2. s_0, s_1 são variáveis aleatórias mutuamente independentes de $\{x_0, \xi_1, \xi_2, \ldots\}$.
3. Existe Ω positivo tal que para cada bola aberta $I \subset \mathrm{B}(\Omega)$, $\mathrm{P}(\xi_t \in I) > 0$ (não ocorrem vazios de probabilidade em $\mathrm{B}(\Omega)$).
4. $\mathrm{E}|x_0| < \infty$.

Condições B2

1. $\gamma(s)$ é uma função monótona decrescente definida em $[0, +\infty)$ pelo que no texto $\gamma(0)$ denota o valor máximo do passo.
2. $\int_0^\infty \gamma(s)ds = \infty$.
3. $\int_0^\infty \gamma^2(s)ds < \infty$.

Condições B3

1. Existe uma função contínua $V(x) : \mathbb{R}^n \to \mathbb{R}^+$ tal que
   (a) $V(x^*) = 0$;
   (b) $\nabla^2 V(x) \leq M$ para cada x, $M > 0$ (o maior valor próprio de $\nabla^2 V(x)$ é inferior a M).
   (c) $\varphi(x)^T \nabla V(x) > 0$ para cada $x \neq x^*$;
   (d) Para cada $\gamma^* < \gamma(0)$ e qualquer z_0, a sequência
   $$z_t = z_{t-1} - \gamma^* \varphi(z_{t-1})$$
   converge, deterministicamente, para x^* e verifica-se que $\{V(z_t), t = 1, 2, \ldots\}$ é monótona decrescente e converge para zero.
2. Existem R e β_0 positivos tais que
   $$\varphi(x)^T \nabla V(x) \geq \frac{1}{2}\gamma(0) \cdot (\varphi(x)^T M \varphi(x) + \mathrm{tr}(S_\xi M)) + \beta_0 \,.$$
   para $|x - x^*| \geq R$. Esta condição limita superiormente o passo máximo $\gamma(0)$ e garante que $\inf_{x \neq x^*} |\varphi(x)| > 0$.

Condições B4

1. u é uma função $\mathbb{R} \to \mathbb{R}$ monótona crescente, para a qual existem limites inferior e superior finitos
   $$\mathrm{u}_+ = \lim_{x \to +\infty} \mathrm{u}(x) > 0 \text{ e } \mathrm{u}_- = \lim_{x \to -\infty} \mathrm{u}(x).$$
2. Denotamos $\mathrm{E}_\omega = \mathrm{E}[\mathrm{u}(X^{(\omega)})]$ onde
   $$X^{(\omega)} = \inf_{\substack{|\varphi_1| \leq \omega \\ |\varphi_2| \leq \omega}} [-(\xi_1 + \varphi_1)^T (\xi_2 + \varphi_2)]$$

Definimos E_0 por $\lim_{\omega\to 0^+}\mathrm{E}_\omega =: \mathrm{E}_0$. Deve verificar-se que E_0 é um número positivo.

A Figura 2.1 na página 17, comum aos casos unidimensional e multidimensional, mostra possíveis exemplos para u onde se incluem algoritmos já estabelecidos.

Comentário 13 *De forma idêntica ao caso unidimensional vamos supor que estamos a observar o processo (2.73), (2.74) começando de $t_0 > 1$. O novo processo com condições iniciais x_{t_0}, s_{t_0}, s_{t_0+1} e a sequência aleatória $\xi_{t_0}, \xi_{t_0+1}, \ldots$ também satisfazem todas as condições. O Lema 15, por exemplo, faz uso deste comentário.*

Comentário 14 *Se* u *ou a distribuição de ξ_t são contínuas, então $\mathrm{E}_0 = \mathrm{E}[\mathrm{u}(-\xi_1^T\xi_2)]$. Mais, se* u *é contínua e satisfaz $\mathrm{u}(x) > -\mathrm{u}(-x)$ quando $x \neq 0$, então B4.2 é válida para qualquer distribuição de ξ_t com variância não nula.*

Comentário 15 *Usamos a notação diferenciada para φ e V: φ' denota uma matriz e ∇V um vector e $\nabla^2 V$ uma matriz.*

Teorema 3 (Cruz, 2005) *Considerando que se verificam as Condições B1 a B4, tem-se, quase-certamente, que $\lim_{t\to\infty} x_t = x^*$.*

As condições para a Normalidade Assimptótica para o campo vectorial, são todas as condições de convergência, às quais acrescentamos as Condições B3.3, B3.4 e B4.3.

Condição B3.3 Todos os valores próprios de $\frac{I}{2} - (1/\mathrm{E}_0)\varphi'(x^*)$ são negativos; I é a matriz identidade.

Condição B3.4 Assume-se que φ se pode decompor em série de Taylor

$$\frac{|\varphi(x) - \varphi'(x^*)\,(x - x^*)|}{|x - x^*|} = \mathrm{o}(1), \text{ quando } x \to x^*\,. \tag{2.75}$$

Comentário 16 *Desta condição segue que*

$$\sup |\varphi(x)|/|x - x^*| < \infty\,. \tag{2.76}$$

pois

$$\frac{|\varphi(x) - \varphi'(x^*)\,(x - x^*)|}{|x - x^*|} \geq \frac{|\varphi(x)|}{|x - x^*|} - |\varphi'(x^*)|$$

e assim

$$\begin{aligned} |o(1)| &\geq \frac{|\varphi(x)|}{|x-x^*|} - |\varphi'(x^*)| \\ \frac{|\varphi(x)|}{|x-x^*|} &\leq |\varphi'(x^*)| - |o(1)| < \infty \end{aligned}$$

Condição B4.3 Assume-se a decomposição de Taylor da função $\mathrm{u}(x+\Delta x) = \mathrm{u}(x) + \mathrm{u}'(\theta)\Delta x$ para θ entre x e $x+\Delta x$.

Teorema 4 (Cruz, 2005) *Seja x_t definido por (2.73) e (2.74) para o qual se supõem verificadas as condições de convergência quase-certa de $x_t \to x^*$. Além destas verificam-se também as Condições B3.3, B3.4 e B4.3, Se $\gamma(s) = 1/s$ (onde $\xrightarrow{d}$ é a convergência em distribuição)*

$$\sqrt{t}(x_t - x^*) \xrightarrow{d} \mathrm{N}(0, V) \tag{2.77}$$

onde a matriz V é definida positiva e é a única solução da equação de Lyapunov (ver Teorema 5 na página 61)

$$\left(\frac{I}{2} - (1/\mathrm{E}_0)\varphi'(x^*)\right)(-V) + (-V)\left(\frac{I}{2} - (1/\mathrm{E}_0)\varphi'(x^*)\right)^T = (1/\mathrm{E}_0)^2 S_\xi\,. \tag{2.78}$$

Comentário 17 *A solução explícita da equação (2.78) é*

$$(-V) = -\int_0^\infty e^{W\cdot t} S e^{W^T\cdot t} dt$$

onde $W = \frac{I}{2} - (1/\mathrm{E}_0)\varphi'(x^*)$, V é definida positiva, e cuja demonstração se encontra, por exemplo, no Teorema 12.3.3 em Lancaster e Tismenetsky [26].

2.5 Demonstração da convergência *quase-certa*

Sem perda de generalidade supomos $x^* = 0$ pelo que $\varphi(x^*) = 0$. A demonstração segue os mesmos passos do caso unidimensional.

Lema 12 *Para cada $\epsilon > 0$ existe $m = m(\epsilon)$ tal que, quase-certamente, ocorre (i) existe t tal que $|x_t| < \epsilon$, ou (ii) existe t tal que $|x_t| < R$ e $s_t \leq m$. (Relembramos que R está definido em B3.2)*

Prova. Fixamos $\epsilon > 0$ e definimos o tempo-de-paragem

$$\tau = \tau(\epsilon, m) = \inf\{t : |x_t| < \epsilon \text{ ou } (|x_t| < R \text{ e } s_t \leq m)\}.$$

O objectivo é provar que para algum m temos $\mathrm{P}(\tau = \infty) = 0$.

Consideramos a sequência $\mathrm{E}_t = \mathrm{E}[V(x_t)\, \mathbb{I}(t < \tau)]$.

Introduzindo as notações simplificadas $V(x_t) = V_t$, $\mathbb{I}(t < \tau) = \mathbb{I}_t$, $\nabla V(x_t) = \nabla_t$, $\gamma(s_t) = \gamma_t$, e usando que $\mathbb{I}_t \leq \mathbb{I}_{t-1}$, obtemos

$$\mathrm{E}_t - \mathrm{E}_{t-1} = \mathrm{E}[V_t I_t - V_{t-1}\, \mathbb{I}_{t-1}] \leq \mathrm{E}[(V_t - V_{t-1})\, \mathbb{I}_{t-1}]. \tag{2.79}$$

De seguinda usamos a decomposição de Taylor

$$V_t = V(x_{t-1} - \gamma_{t-1} y_t) = V_{t-1} - \gamma_{t-1} y_t^T \nabla_{t-1} + \frac{1}{2}\gamma_{t-1}^2 y_t^T \nabla^2 V_{t-1}(x') y_t,$$

em que x' é um ponto entre x_t e x_{t-1}. Substituindo $y_t = \varphi_{t-1} + \xi_t$ e, de acordo com B3.1, obtemos

$$V_t - V_{t-1} \leq -\gamma_{t-1}\varphi_{t-1}^T \nabla_{t-1} - \gamma_{t-1}\xi_t^T \nabla_{t-1} + \frac{1}{2}\gamma_{t-1}^2(\varphi_{t-1}^T M \varphi_{t-1} + \xi_t^T M \xi_t). \tag{2.80}$$

Usando (2.79) e (2.80) e tomando em conta que cada um dos valores γ_{t-1}, φ_{t-1}, $\mathbb{I}_{t-1}$ é determinado por x_{t-1} e s_{t-1} e portanto mutuamente independentes de ξ_t (Condição B1.1), temos

$$\begin{aligned}
&\mathrm{E}_t - \mathrm{E}_{t-1} \leq \\
&\quad \leq \quad \mathrm{E}[-\gamma_{t-1}\varphi_{t-1}^T \nabla_{t-1} - \gamma_{t-1}\xi_t^T \nabla_{t-1} + \frac{1}{2}\gamma_{t-1}^2(\varphi_{t-1}^T M \varphi_{t-1} + \xi_t^T M \xi_t)\, \mathbb{I}_{t-1}] = \\
&\quad = \quad \mathrm{E}[-\gamma_{t-1}\varphi_{t-1}^T \nabla_{t-1}] + \\
&\qquad \mathrm{E}[-\gamma_{t-1}\xi_t^T \nabla_{t-1}] + \\
&\qquad \mathrm{E}[\frac{1}{2}\gamma_{t-1}^2(\varphi_{t-1}^T M \varphi_{t-1})\, \mathbb{I}_{t-1}] + \\
&\qquad \mathrm{E}[\frac{1}{2}\gamma_{t-1}^2\, \mathbb{I}_{t-1}] \cdot \mathrm{E}[\xi_t^T M \xi_t]
\end{aligned}$$

e esperança de v.a. independentes temos

- $\mathrm{E}[-\gamma_{t-1}\xi_t^T \nabla_{t-1}] = 0$;
- $\mathrm{E}[\xi_t^T M \xi_t] \leq \mathrm{tr}(S_\xi M)$

e então

$$\begin{aligned}
&\mathrm{E}_t - \mathrm{E}_{t-1} \leq \\
&\quad \leq \quad \mathrm{E}[-\varphi_{t-1}^T \nabla_{t-1} + \frac{1}{2}\gamma_{t-1}(\varphi_{t-1}^T M \varphi_{t-1} + \mathrm{tr}(S_\xi M)))\gamma_{t-1}\, \mathbb{I}_{t-1}]\,.
\end{aligned} \tag{2.81}$$

Se $\mathbb{I}_{t-1} = 1$, então temos *(i)* $|x_t| \geq R$, ou *(ii)* $|x_t| \geq \epsilon$ e $s_t \geq m$. No caso *(i)* usando B3.2 obtemos

$$-\varphi_{t-1}^T \nabla_{t-1} + \frac{1}{2}\gamma_{t-1}(\varphi_{t-1}^T M \varphi_{t-1} + \mathrm{tr}(S_\xi M)) \leq -\beta_0 \,. \tag{2.82}$$

No caso *(ii)* é válido $\gamma_t < \gamma(m)$ e definimos $\delta_\epsilon := \inf\{\varphi(x)^T \nabla V(x), \text{ para todo } |x| \geq \epsilon\}$. Neste contexto temos

$$-\varphi_{t-1}^T \nabla_{t-1} + \frac{1}{2}\gamma_{t-1}(\varphi_{t-1}^T M \varphi_{t-1} + \mathrm{tr}(S_\xi M)) \leq \tag{2.83}$$

$$\leq \quad -\delta_\epsilon + \frac{1}{2}\gamma(m)(\varphi_{t-1}^T M \varphi_{t-1} + \mathrm{tr}(S_\xi M)) := -\beta(\epsilon, m) \tag{2.84}$$

Escolhemos m tal que $\beta(\epsilon, m) > 0$ e denotamos $\beta = \inf\{\beta_0, \beta(\epsilon, m)\}$. Assim, em ambos os casos, a expressão entre parentesis rectos no L.D. de (2.81) é inferior a $-\beta \cdot \gamma_{t-1}\, \mathbb{I}_{t-1}$ e por isso

$$\mathrm{E}_t - \mathrm{E}_{t-1} \leq -\beta \cdot \mathrm{E}[\gamma_{t-1}\, \mathbb{I}_{t-1}].$$

Usando que $s_t \leq s_0 + t\mathrm{u}_+$ e $\mathrm{E}\,\mathbb{I}_t = \mathrm{P}(t < \tau)$ temos

$$\mathrm{E}_t - E_{t-1} \leq -\beta\, \gamma(s_0 + t\mathrm{u}_+)\, \mathrm{P}(t < \tau),$$

e porque $\mathrm{P}(j < \tau) \geq \mathrm{P}(t < \tau)$ quando $j < t$ temos, aplicando o argumento da indução, que

$$E_t \leq E_1 - \beta \mathrm{P}(t < \tau) \sum_{j=0}^{t-1} \gamma(s_0 + j\mathrm{u}_+) \,.$$

onde $\tilde{E}_0 := E(V(x_0)\, \mathbb{I}(0 < \nu)) < \infty$ pela Condição B1.4.

A função V é positiva para $x \neq x^*$, portanto $\tilde{E}_t \geq 0$, e daqui segue

$$\mathrm{P}(t < \tau) < \frac{\tilde{E}_0}{\beta \sum_{j=0}^{t-1} \gamma(s_0 + j\mathrm{u}_+)}.$$

Quando $t \to \infty$ e usando $\sum_{j=0}^{\infty} \gamma(s_0 + j\mathrm{u}_+) = \infty$ (inferido da Condição B2.2), podemos concluir que $P(\tau = \infty) = 0$.

□

Lema 13 *Para cada $\epsilon > 0$ e $m > 0$ existe δ positivo tal que se $|x_0| < R$ e $s_0 \leq m$ então*

$$\mathrm{P}(\textit{existe } t, |x_t| < \epsilon) \geq \delta \,.$$

Prova. Consideramos a função V definida nas Condições B4. Sejam

$$\bar{\epsilon} = \inf\{V(x), |x| \geq \epsilon\}$$

$$\bar{R} = \sup\{V(x), |x| \leq R\}$$

então $|x_0| \leq R \Rightarrow V(x_0) \leq \bar{R}$ e $V(x) < \bar{\epsilon} \Rightarrow |x| < \epsilon$.

Vamos mostrar que $V(x_t) < \bar{\epsilon}$ para algum t. Denotamos $V_t := V(x_t)$ e reescrevemos

$$V_t = V_0 \frac{V_1}{V_0} \frac{V_2}{V_1} \cdots \frac{V_t}{V_{t-1}}.$$

Definimos o processo determinístico de passo constante $\rho \leq \gamma(0)$

$$z_t = z_{t-1} - \rho\varphi(z_{t-1}), \quad t = 1, 2, \ldots$$

e pela Condição B3.1, existe $V(\cdot)$ tal que $\{V(z_t)\}$ converge monotonamente para zero. Usando o desenvolvimento de Taylor

$$\begin{aligned} V(z_t) &= V(z_{t-1} - \rho\varphi(z_{t-1})) = \\ &= V(z_{t-1}) - \rho\varphi(z_{t-1})^T \nabla V(z_{t-1}) + \\ &\quad + \frac{\rho^2}{2} \varphi(z_{t-1})^T \nabla^2 V(z') \varphi(z_{t-1}) \\ &= V(z_{t-1}) - \rho \times \\ &\quad (\varphi(z_{t-1})^T \nabla V(z_{t-1}) - \frac{\rho}{2} \varphi(z_{t-1})^T \nabla^2 V(z') \varphi(z_{t-1})) \end{aligned}$$

para um certo vector z' entre z_t e z_{t-1}. Definimos

$$U(z, \rho) := \frac{1}{V(z)} \times \left(\varphi(z)^T \nabla V(z) - \frac{\rho}{2} \varphi(z)^T \nabla^2 V(z') \varphi(z) \right)$$

em que z' está entre z e $z - \rho\varphi(z)$ e como $V(z_t)$ decresce monotonamente, é necessário que $U(\cdot, \cdot) > 0$. Definimos

$$\bar{U} := \inf_{\substack{\epsilon \leq |z| \leq R \\ \rho \leq \gamma(0)}} U(z, \rho)$$

onde $\bar{U}$ é uma constante positiva porque $U(\cdot, \cdot) > 0$ nos intervalos $\epsilon \leq |z| \leq R$ e $\rho \leq \gamma(0)$.

Consideramos a expansão de Taylor ao processo original

$$\begin{aligned} V(x_t) &= V(x_{t-1} - \gamma(s_{t-1})\varphi(x_{t-1}) - \gamma(s_{t-1})\xi_t)) \\ &= V(x_{t-1} - \gamma(s_{t-1})\varphi(x_{t-1})) - \\ &\quad - \gamma(s_{t-1}) \xi_t^T \nabla V(x_{t-1} - \gamma(s_{t-1})\varphi(x_{t-1})) + \frac{\gamma(s_{t-1})}{2} \xi_t^T \nabla V^2(x'') \xi_t \end{aligned}$$

e fazendo $\zeta_t := |\xi_t|$ temos para a última parcela

$$\begin{aligned} - \gamma(s_{t-1}) \xi_t^T \nabla V(x_{t-1} - \gamma(s_{t-1})\varphi(x_{t-1})) + \frac{\gamma^2(s_{t-1})}{2} \xi_t^T \nabla^2 V(x'') \xi_t &\leq \\ \gamma(0) \zeta_t |\nabla V(x_{t-1} - \gamma(s_{t-1})\varphi(x_{t-1}))| + \frac{\gamma^2(0)}{2} \zeta_t^2 M &\leq \\ \zeta_t C_\xi \end{aligned}$$

com as seguintes justificações:

1. impondo que $\zeta_t < 1$;

2. dado que $\epsilon \leq |x| \leq R$ então x_{t-1} e $\varphi(x_{t-1})$ são vectores dum conjunto fechado e limitado e $\gamma(s_{t-1}) \leq \gamma(0)$ e assim $\nabla V(x_{t-1} - \gamma(s_{t-1})\varphi(x_{t-1}))$ pode ser majorado por uma constante.

Assim, recordando a função $U(\cdot,\cdot)$,

$$V(x_t) \leq V(x_{t-1})(1 - \gamma(s_{t-1}) \cdot U(x_{t-1}, \gamma(s_{t-1}))) + \zeta_t \cdot C_\xi \,.$$

Usando que $1/V(x) \leq 1/\bar{\epsilon}$, para $\epsilon \leq |x| \leq R$, e que $\gamma(s_{t-1}) > \gamma(m + (t-1) \cdot u_+)$,

$$\begin{aligned} \frac{V_t}{V_{t-1}} &= 1 - \gamma(s_{t-1}) \cdot \bar{U} + \zeta_t \cdot C_\xi/\bar{\epsilon} \\ &\leq 1 - \gamma(m + (t-1)u_+) \cdot \bar{U} + \zeta_t \cdot C_\xi/\bar{\epsilon} \end{aligned}$$

Fazendo $G_t := 1 - \gamma(m + (t-1)u_+) \cdot \bar{U}$ se tem $G_t < 1$. A divergência da série $\sum_t \gamma(m + t \cdot u_+)$ implica que o produtório $\prod_{i=1}^{t-1} G_i$ tende para zero. Usando que $G_t \leq \sqrt{G_t} < 1$ podemos escolher ζ_t tal que

$$G_t + \zeta_t \cdot C_\xi/\bar{\epsilon} \leq \sqrt{G_t} < 1 \tag{2.85}$$

e

$$\frac{V_t}{V_{t-1}} \leq \sqrt{G_t}$$

sempre que $\epsilon \leq |x_{t-1}| \leq R$ e $|\xi_t| < \zeta_t < 1$. Escolhemos n tal que $\bar{R} \prod_{i=1}^{n-1} \sqrt{G_t} < \bar{\epsilon}$ e supondo que $|x_0| < R$, $s_0 \leq m$ e $|\xi_t| < \zeta_t$ enquanto $1 \leq t \leq n-1$. Então, para algum $t \in \{1, \ldots, n\}$, $|x_t| < \epsilon$ com probabilidade superior a

$$\delta := \mathrm{P}(|\xi_1| < \zeta_1, |\xi_2| < \zeta_2, \ldots, |\xi_n| < \zeta_n)$$

□

Dos Lemas 12 e 13 vemos que para cada $\epsilon > 0$ existe $\delta > 0$ tal que para condições iniciais arbitrárias x_0, s_0, s_1

$$\mathrm{P}(\text{para alguns } t, |x_t| < \epsilon) > \delta.$$

Então podemos escolher um número inteiro positivo $n = n(x_0, s_0, s_1)$ tal que

$$\mathrm{P}(\text{para alguns } t \leq n, |x_t| < \epsilon) > \delta/2 \,.$$

Denotamos $\bar{p} = \sup \mathrm{P}(\text{para cada } t, |x_t| \geq \epsilon)$, sendo o supremo tomado sobre todas as condições iniciais x_0, s_0, s_1. Fixamos x_0, s_0, s_1; então

$$\begin{aligned}
&\mathrm{P}(\text{para cada } t, |x_t| \geq \epsilon) = \\
&= \mathrm{P}(\text{para cada } t > n, |x_t| \geq \epsilon \,\Big|\, \text{para cada } t \leq n, |x_t| \geq \epsilon) \cdot \mathrm{P}(\text{para cada } t \leq n, |x_t| \geq \epsilon) \leq \\
&\leq \bar{p}\,(1 - \delta/2). \qquad (2.86)
\end{aligned}$$

O supremo do L.E. de (2.86) sobre todos os triplos (x_0, s_0, s_1) é $\bar{p}$. Assim obtemos a desigualdade $\bar{p} \leq \bar{p}\,(1 - \delta/2)$ de onde $\bar{p} = 0$. Obtemos pois o seguinte Lema:

Lema 14 *Para cada $\epsilon > 0$, quase-certamente existe t tal que $|x_t| < \epsilon$.*

Lema 15 *Fixamos $\epsilon > 0$ e $\eta > 0$. Então existe $\epsilon_1 > 0$ e $\delta > 0$ tal que se $|x_0| < \epsilon_1$ então*

$$\mathrm{P}(\textit{para alguns } t,\ |x_t| < \epsilon \textit{ e } s_t \geq \eta) > \delta$$

Prova. Partindo de $x_t = x_0 - \sum_{i=1}^{t} \gamma_{i-1} y_i$ e usando o desenvolvimento de Taylor,

$$\begin{aligned}
V(x_t) &= V(x_0 - \sum_{i=1}^{t} \gamma_{i-1} y_i) \leq \\
&\leq V(x_0) + |\nabla V(x_0)| \sum_{i=1}^{t} \gamma_{i-1} |y_i| \cos(y_i, \nabla V(x_0)) + C_1 |\sum_{i=1}^{t} \gamma_{i-1} y_i|^2 .
\end{aligned}$$

Para garantir o aumento do contador do passo s_t requerido pelo Lema consideramos duas secções cónicas simétricas onde permanecerão os vectores y_t e onde definimos uma amplitude máxima e mínima para $|y_t|$, $y_I \leq |y_t| \leq y_{II}$, com y_I, y_{II} a definir. Tomamos como referência o ponto x_0 e o gradiente nesse ponto $\nabla_0 := \nabla V(x_0)$. Como se verá adiante interessa limitar o produto interno

$$y^T \nabla V(x_0) = |y_t| \cdot |\nabla_0| \cdot \cos(y_t, \nabla_0)$$

Escolhemos que $y_{\text{ímpar}}$ pertence à secção cónica no lado oposto a ∇_0 e y_{par} à outra secção. Fixamos um valor θ para o ângulo interno do cone centrado no vector ∇_0 e onde θ pertence a $(0, \pi/2)$. Nesta situação o $\cos(y_t, \nabla_0)$ é enquadrado por

$$-1 \leq \cos(y_t, \nabla_0) \leq -\cos(\theta), \quad t \text{ ímpar}, \qquad (2.87)$$

$$\cos(\theta) \leq \cos(y_t, \nabla_0) \leq 1, \quad t \text{ par}\,. \qquad (2.88)$$

Usando (2.87) e (2.88) temos

$$-y_{II} \leq |y_t| \cos(y_1, \nabla_0) \leq -y_I \cos(\theta), \quad \text{caso ímpar,} \tag{2.89}$$

$$y_I \cos(\theta) \leq |y_t| \cos(y_2, \nabla_0) \leq y_{II}, \quad \text{caso par.} \tag{2.90}$$

É possível garantir $V(x_t) < \bar{\epsilon}$ se mostrarmos que

$$V(x_0) < \bar{\epsilon}/3 \tag{2.91}$$

$$\left| \sum_{i=1}^{t} \gamma_{i-1} |y_i| |\nabla_0| \cos(y_i, \nabla_0) \right| < \bar{\epsilon}/3 \tag{2.92}$$

$$C_1 |\sum_{i=1}^{t} \gamma_{i-1} y_i|^2 < \bar{\epsilon}/3 \tag{2.93}$$

De (2.91) é possível determinar ϵ_1 pela Condição B3.3.

De (2.93) concluímos

$$C_1 |\sum_{i=1}^{t} \gamma_{i-1} y_i|^2 \leq C_1 y_{II}^2 \sum_{i=1}^{\infty} \gamma_{i-1}^2 < \bar{\epsilon}/3 \tag{2.94}$$

de onde podemos escolher y_{II}, porque pela Condição B2.2 a série é convergente.

Como y_t pertence às secções cónicas de forma alternada temos

$$\mathrm{u}(-y_t^T y_{t-1}) \leq \mathrm{u}(y_I^2 \cos(\pi - \theta)) = \mathrm{u}(-y_I^2 \cos\theta), \quad t = 1, 2, \ldots, n-1$$

de onde

$$s_t \geq (t-2)\mathrm{u}(-y_I^2 \cos\theta), \quad t = 3, 4, \ldots, n \tag{2.95}$$

Para satisfazer $s_t \geq \eta$ requerido pelo Lema, assumimos $y_I \geq y_{II}/2$ e temos

$$n - 2 \geq \frac{\eta}{\mathrm{u}(-(y_{II}^2/4)\cos\theta)} \tag{2.96}$$

obtido de (2.95).

Desenvolvendo o L.E. de (2.92) temos por (2.89) e (2.90)

$$\begin{aligned} -y_{II} \sum_{\substack{i=1 \\ (\text{ímpar})}}^{t} \gamma_{i-1} + y_I \cos(\theta) \sum_{\substack{i=1 \\ (\text{par})}}^{t} \gamma_{i-1} \leq \\ \leq \sum_{i=1}^{t} \gamma_i |y_i| \, |\nabla_0| \, \cos(y_i, \nabla_0) \leq \\ \leq -y_I \cos(\theta) \sum_{\substack{i=1 \\ (\text{ímpar})}}^{t} \gamma_{i-1} + y_{II} \sum_{\substack{i=1 \\ (\text{par})}}^{t} \gamma_{i-1} \, . \end{aligned} \tag{2.97}$$

A soma ímpar é maior que a par começando em $i = 1$. Temos

$$\left| \sum_{i=1}^{t} \gamma_{i-1} |y_i| \, |\nabla_0| \, \cos(y_i, \nabla_0) \right| \leq y_{II} \sum_{\substack{i=1 \\ (\text{ímpar})}}^{t} \gamma_{i-1} - y_I \cos(\theta) \sum_{\substack{i=1 \\ (\text{par})}}^{t} \gamma_{i-1} \tag{2.98}$$

A condição (2.92), com auxílio de (2.98), é satisfeita se

$$y_{II} \sum_{\substack{i=1 \\ (\text{ímpar})}}^{t} \gamma_{i-1} - y_I \cos(\theta) \sum_{\substack{i=1 \\ (\text{par})}}^{t} \gamma_{i-1} \leq \bar{\epsilon}/3 \tag{2.99}$$

onde verificamos que, fazendo $t := n$ em (2.99), e uma vez fixo o valor da primeira parcela por y_{II} já dado, podemos escolher y_I tão grande quanto se deseje, satisfazendo o majorante desejado.

Para determinar δ temos em conta que em cada iteração t o valor de $\varphi(x_t) := \varphi_t$, y_I, y_{II}, θ são conhecidos. Fazendo

$$v_t := \frac{(\varphi_{t-1} + \xi_t)^T \nabla_0}{|y_t| \cdot |\nabla_0|}$$

as condições que determinam uma região admissível para cada vector aleatório ξ_t resumem-se a

$$\begin{aligned} & y_I \leq |\varphi_{t-1} + \xi_t| \leq y_{II} \\ & \pi \leq \cos^{-1}(v_t) \leq \pi - \theta, \quad t \text{ ímpar} \\ & 0 \leq \cos^{-1}(v_t) \leq \theta, \quad t \text{ par} \end{aligned} \tag{2.100}$$

Definimos δ_1 como sendo a menor probabilidade das regiões definidas em cada iteração $t = 1, \ldots, n$ e definimos que $\delta := \delta_1^n$, terminando a demonstração (δ_1 é um número positivo pela Condição B1.3).

□

Dos Lemas 14 e 15 segue que para cada $\epsilon > 0$ e $\eta > 0$ a probabilidade de que para alguns t, $|x_t| < \epsilon$ e $s_t \geq \eta$ seja maior que $\delta > 0$, depende somente de ϵ e η. Repetindo o argumento do Lema 14 temos

Lema 16 *Para cada $\epsilon > 0$ e $\eta > 0$, quase-certamente existe t tal que $|x_t| < \epsilon$ e $s_t \geq \eta$.*

Definimos o tempo-de-paragem $\tau(\epsilon) = \inf\{t : |x_t| \geq \epsilon\}$.

Lema 17 *Para cada $0 < \theta < \mathrm{E}_0$ existe uma constante $\epsilon_0 > 0$ e uma sequência π_n tal que $\lim_{n\to\infty} \pi_n = 0$ e*

$$\mathrm{P}(s_t > s_0 + t\theta - n \text{ para cada } t < \tau(\epsilon_0)) > 1 - \pi_n.$$

Prova. Vamos mostrar que

$$\mathrm{P}(\text{existe } t < \tau(\epsilon_0) \text{ tal que } s_t \leq s_0 + t\theta - n) \leq \pi_n \to 0\,.$$

De B4.2 segue que para algum ω_0 positivo existe $\mathrm{E}_{\omega_0} > \theta$ onde $\mathrm{E}_{\omega_0} = \mathrm{E}[u(X^{(\omega_0)})]$ e

$$X^{(\omega_0)} = \inf_{\substack{|\varphi_1|\leq\omega_0 \\ |\varphi_2|\leq\omega_0}} [-(\xi_1+\varphi_1)^T(\xi_2+\varphi_2)]. \tag{2.101}$$

Escolhemos ϵ_0 tal que

$$\sup_{|x|<\epsilon_0} |\varphi(x)| \leq \omega_0.$$

Definimos a sequência $\{\tilde{s}_t\}$ por

$$\tilde{s}_0 = s_0; \quad \tilde{s}_t = \tilde{s}_{t-1} + \mathrm{u}(X_t^{(\omega_0)}) \tag{2.102}$$

onde

$$X_t^{(\omega_0)} = \inf_{\substack{|\varphi_{t-1}|\leq\omega_0 \\ |\varphi_{t-2}|\leq\omega_0}} [-(\xi_t+\varphi_{t-1})^T(\xi_{t-1}+\varphi_{t-2})]. \tag{2.103}$$

Comparando (2.102) e (2.103) com (2.74), para $t < \tau(\epsilon_0)$, obtemos

$$\tilde{s}_t \leq s_t. \tag{2.104}$$

De (2.102) segue que

$$\tilde{s}_t - s_0 = t\mathrm{E}_{\omega_0} + \mathbb{I}_t^{\mathrm{par}} + \mathbb{I}_t^{\mathrm{ímpar}} \tag{2.105}$$

sendo

$$\mathbb{I}_t^{\mathrm{par}} = \sum_{\substack{i=1 \\ (i\,\mathrm{par})}}^{t} [\mathrm{u}(X_t^{(\omega_0)}) - \mathrm{E}_{\omega_0}], \quad \mathbb{I}_t^{\mathrm{ímpar}} = \sum_{\substack{i=1 \\ (i\,\mathrm{ímpar})}}^{t} [\mathrm{u}(X_t^{(\omega_0)}) - \mathrm{E}_{\omega_0}]$$

em que $\mathbb{I}_t^{\mathrm{par}}$ e $\mathbb{I}_t^{\mathrm{ímpar}}$ são somas de variáveis limitadas, independentes e identicamente distribuídas com média zero e variância linear com t.

Comentário 18 *Aqui repetimos o argumento do caso unidimensional. Apesar de assimptoticamente normais, $\mathbb{I}_t^{par}$ e $\mathbb{I}_t^{ímpar}$ são dependentes entre si pelo que usamos o seguinte argumento para estimar a probabilidade da sua soma: O argumento é: $X + Y < a$ implica que $X < a/2$ ou $Y < a/2$ onde X e Y são variáveis aleatórias reais e a uma constante real. Assim*

$$\mathrm{P}(X + Y < a) \leq \mathrm{P}(X < a/2) + \mathrm{P}(Y < a/2) \simeq 2\mathrm{P}(X < a/2).$$

Então, por Var $\mathbb{I}_t^{\text{par}} = t \cdot V_{\mathbb{I}_1}$, temos

$$\mathrm{P}(\ \mathbb{I}_t^{\text{par}} + \ \mathbb{I}_t^{\text{ímpar}} < 2a) \lesssim 2\mathrm{P}(\ \mathbb{I}_t^{\text{par}} < a) \leq 2\Phi(\frac{a}{\sqrt{t}\sqrt{V}_{\mathbb{I}_1}}). \tag{2.106}$$

Do evento $s_t \leq s_0 + t\theta - n$, sabendo que $\tilde{s}_t \leq s_t$ para $t < \tau(\epsilon_0)$, segue

$$\begin{aligned} \tilde{s}_t &\leq s_0 + t\theta - n \Leftrightarrow \\ s_0 + t\mathrm{E}_{\omega_0} + \ \mathbb{I}_t^{\text{par}} + \ \mathbb{I}_t^{\text{ímpar}} &\leq s_0 + t\theta - n \Leftrightarrow \\ \mathbb{I}_t^{\text{par}} + \ \mathbb{I}_t^{\text{ímpar}} &\leq -t(\mathrm{E}_{\omega_0} - \theta) - n \end{aligned} \tag{2.107}$$

Comentário 19 *No que se segue usamos a seguinte expressão, onde* $\{X_i, i = 1, \ldots, \infty\}$ *é uma sequência de variáveis aleatórias,*

$$\mathrm{P}(\textit{existe } t < \tau \textit{ tal que } X_t < a) \leq \sum_{i=1}^{\tau} \mathrm{P}(X_i < a) \leq \sum_{i=1}^{\infty} \mathrm{P}(X_i < a) \tag{2.108}$$

Por (2.106), (2.107) e (2.108) segue

$$\begin{aligned} \mathrm{P}(\text{existe } t < \tau(\epsilon_0) \text{ tal que } s_t \leq s_0 + t\theta - n) &\leq \\ \mathrm{P}(\text{existe } t < \tau(\epsilon_0) \text{ tal que } \ \mathbb{I}_t^{\text{par}} + \ \mathbb{I}_t^{\text{ímpar}} \leq -t(\mathrm{E}_{\omega_0} - \theta) - n) &\leq \\ \sum_{i=1}^{\infty} \mathrm{P}(\ \mathbb{I}_i^{\text{par}} + \ \mathbb{I}_i^{\text{ímpar}} \leq -i(\mathrm{E}_{\omega_0} - \theta) - n) &\lesssim \\ 2\sum_{i=1}^{\infty} \mathrm{P}(\frac{\mathbb{I}_i^{\text{par}}}{\sqrt{i}\sqrt{V_I}} \leq -\sqrt{i}\frac{\mathrm{E}_{\omega_0} - \theta}{\sqrt{V}_I} - \frac{n}{\sqrt{i}\sqrt{V_I}}) &\leq \\ 2\sum_{i=1}^{\infty} \Phi(-\sqrt{i}K_1 - \frac{n}{\sqrt{i}}K_2) &:= \pi_n \end{aligned}$$

para certas constantes $K_1 > 0$ e $K_2 > 0$. A série da última desigualdade é convergente pelo que $\pi_n \to 0$ e então

$$\begin{aligned} \pi_n := \mathrm{P}(&\text{existe } t \text{ tal que } \ \mathbb{I}_t^{\text{par}} + \ \mathbb{I}_t^{\text{ímpar}} \leq \\ &\leq -n - t(\mathrm{E}_{\omega_0} - \theta)) \to 0 \text{ quando } n \to \infty. \end{aligned}$$

□

Fixamos θ e ϵ_0 como no Lema 17, e escolhemos arbitrariamente valores positivos $\epsilon < \epsilon_0$ e n. Definimos o tempo-de-paragem

$$\nu = \nu(n, \epsilon) = \inf\{t : |x_t| \geq \epsilon \text{ ou } s_t \leq s_0 - n + t\theta\}$$

e escolhemos $\epsilon_1 > 0$ tal que

$$\sup_{|x| < \epsilon_1} V(x) < \frac{1}{2} \inf_{|x| > \epsilon} V(x).$$

Lema 18 *Seja* $|x_0| < \epsilon_1$, *então*

$$\mathrm{P}(\nu < \infty) \leq K \int_{s_0-n-1}^{\infty} \gamma^2(s)ds + \pi_n,$$

sendo K *uma constante que depende de* ϵ.

Prova. Usando (2.80) no Lema 12, temos

$$V_t - V_{t-1} \leq -\gamma_{t-1}\varphi_{t-1}^T \nabla V_{t-1} - \gamma_{t-1}\xi_t^T \nabla V_{t-1} + 1/2\gamma_{t-1}^2(\varphi_{t-1}^T M\varphi_{t-1} + \xi_t^T M\xi_t)$$

e fazemos $V_t - V_0 \leq I_t' + I_t''$ onde

$$I_t' = \left| \sum_{i=1}^{t} \gamma_{i-1}\varphi_{i-1}^T \nabla V_{i-1} + \gamma_{i-1}\xi_i^T \nabla V_{i-1} \right|$$
$$I_t'' = 1/2 \sum_{i=1}^{t} \gamma_{i-1}^2(\varphi_{i-1}^T M\varphi_{i-1} + \xi_i^T M\xi_i).$$

Seja $\delta := (1/2)\inf_{|x|>\epsilon} V(x)$. Como $|x_t| > \epsilon$ então $V_t - V_0 > \delta$, ou seja,

$$I_t' + I_t'' \geq V_t - V_0 > \delta,$$

o que implica $I_t' > \delta/2$ ou $I_t'' > \delta/2$. Com o objectivo de estimar $\mathrm{P}(\nu < \infty)$ denotamos

$$P' = \mathrm{P}(I_\nu' \, \mathbb{I}(\nu < \infty) > \delta/2)$$
$$P'' = \mathrm{P}(I_\nu'' \, \mathbb{I}(\nu < \infty) > \delta/2)$$

e usando o Lema 17 temos

$$\mathrm{P}(\nu < \epsilon) \leq \pi_n + P' + P''. \tag{2.109}$$

Pela desigualdade de Markov (por exemplo, [48, p. 59]), $\mathbb{I}^2(\cdot) = \mathbb{I}(\cdot)$, e $\mathbb{I}(i-1 < \nu < \infty) < \mathbb{I}(i-1 < \nu)$, temos

$$\begin{aligned} P' &\leq \frac{4}{\delta^2}\mathrm{E}[{I_\nu'}^2 \, \mathbb{I}^2(\nu < \infty)] = \\ &= \frac{4}{\delta^2}\mathrm{E}\left[\left(\sum_{i=1}^{\nu-1} \gamma_{i-1}(\varphi_{i-1}^T + \xi_i^T)\nabla V_{i-1})\right)^2 \cdot \mathbb{I}(\nu < \infty)\right] \\ &= \frac{4}{\delta^2}\sum_{i,j=1}^{\infty} \mathrm{E}[\gamma_{i-1}(\varphi_{i-1}^T + \xi_i^T)\nabla V_{i-1} \, \mathbb{I}(i-1 < \nu) \times \\ &\qquad \times \gamma_{j-1}(\varphi_{j-1}^T + \xi_j^T)\nabla V_{j-1} \, \mathbb{I}(j-1 < \nu)]. \end{aligned}$$

Relembrando que as variáveis γ_{i-1}, V_{i-1}, $\mathbb{I}(i-1 < \nu)$ e ξ_i são mutuamente independentes, concluímos que as parcelas em que $i \neq j$ valem zero. Então

$$P' \leq \frac{4}{\delta^2}\sum_{i=1}^{\infty} \mathrm{E}[\gamma_{i-1}^2(\varphi_{i-1}^T \nabla V_{i-1})^2(\xi_i^T \nabla V_{i-1})^2 \, \mathbb{I}(i-1 < \nu)] \leq K'\mathrm{E}\sum_{i=1}^{\nu-1} \gamma_{i-1}^2 \tag{2.110}$$

em que K' é uma constante que verifica

$$(4/\delta^2)\cdot \sup_{|x|<\epsilon}(\varphi_{i-1}^T\nabla V_{i-1})^2\cdot \sup_{|x|<\epsilon}\mathrm{E}[\xi_i^T\nabla V_{i-1}]^2 < K'.$$

Por $\mathrm{P}(X>\delta/2)\le \frac{\mathrm{E}|X|}{2/\delta}$ temos

$$P'' \le \frac{2}{\delta}(1/2)\mathrm{E}[\sum_{i=1}^{\nu-1}\gamma_{i-1}^2(\varphi_{i-1}^T M\varphi_{i-1}+\xi_i^T M\xi_i)] \le K''\sum_{i=1}^{\nu-1}\gamma_{i-1}^2 \tag{2.111}$$

em que K'' verifica

$$(2/\delta)\sup_{|x|<\epsilon}\varphi_{t-1}^T M\varphi_i + \mathrm{E}\xi_i^T M\xi_i < K''$$

porque $\mathrm{E}\xi\xi^T := S_\xi$.

Para $t<\nu$, $s_t > s_0+t\theta-n$, então $\gamma_t<\gamma(s_0-n+t\theta)$, e

$$\mathrm{E}\left[\sum_{i=1}^{\nu-1}\gamma_i^2\right] < \sum_{i=1}^{\infty}\gamma^2(s_0-n+i\theta) \le \frac{1}{\theta}\int_{s_0-n-1}^{\infty}\gamma^2(s)ds. \tag{2.112}$$

Tomando $K=\theta^{-1}(K'+K'')$, de (2.109), (2.110), (2.111) e (2.112) obtemos o Lema 18. □

Agora, fixamos $\epsilon<\epsilon_0$ positivo e escolhemos n e η tal que $1-\pi_n-K\int_{\eta-n-1}^{\infty}\gamma^2(s)ds =: \delta$ seja positivo. Fixamos também $\epsilon_1=\epsilon_1(\epsilon)$ como definido acima. De acordo com os Lemas 16 e 18, *quase-certamente* existe t_0 tal que $|x_{t_0}|<\epsilon_1$, $s_{t_0}\ge\eta$, e a probabilidade para todo o $t\ge t_0$, $|x_t|<\epsilon$ excede δ.

Definimos a sequência de tempos-de-paragem $\tau_1=1$,

$$\tau_{i+1}=\inf\{\tau>\tau_i : |x_\tau|\ge\epsilon, \text{ e para algum } \tau_i\le t<\tau,\ |x_t|<\epsilon_1 \text{ e } s_t>\eta\},\quad i=1,2,\ldots .$$

Temos

$$\mathrm{P}(\tau_{i+1}=\infty \,|\, \tau_i<\infty)\ge\delta,$$

de onde

$$\mathrm{P}(\tau_{i+1}<\infty)=\mathrm{P}(\tau_{i+1}<\infty\,|\,\tau_i<\infty)\,\mathrm{P}(\tau_i<\infty)\le(1-\delta)\,\mathrm{P}(\tau_i<\infty).$$

Assim, $\mathrm{P}(\tau_i<\infty)\to 0$ quando $i\to\infty$; isto implica que *quase-certamente* $i_0=\sup\{i:\tau_i<\infty\}$ é finito.

De acordo com o Lema 16, *quase-certamente* existe $t_0\ge\tau_{i_0}$ tal que $|x_{t_0}|<\epsilon_1$ e $s_{t_0}>\eta$; daqui concluímos que $|x_t|<\epsilon$ quando $t>t_0$. O Teorema 3 está demonstrado. □

2.6 Demonstração da normalidade assimptótica

O corpo central da demonstração da normalidade assimptótica segue o trabalho de Delyon and Juditsky [14] e será apresentado depois do próximo Lema que revela o comportamento determinístico do passo $\gamma(s_t) := 1/s_t$.

Recordamos a definição de E_0 na Condição B4.2.

Lema 19 *Sejam s_0 e s_1 variáveis aleatórias que são condição inicial do processo $\{s_t\}$, definido em (2.74). Então*

$$\gamma(s_t) = 1/s_t = \frac{1}{\mathrm{E}_0 t}(1 + \mathrm{o}_t), \text{ quase-certamente} \tag{2.113}$$

onde o_t é uma v.a. que depende de $\{\xi_i, i \leq t\}$ e das condições iniciais e que verifica $\lim_{t\to\infty} \mathrm{o}_t = 0$ quase-certamente.

Prova. A Condição B4.3 permite a decomposição

$$\begin{aligned}
\mathrm{u}(-y_{i-1}^T y_i) &= \mathrm{u}(-(\varphi_{i-2} + \xi_{i-1})^T(\varphi_{i-1} + \xi_i)) = \\
&= \mathrm{u}(-(\varphi_{i-2} + \xi_{i-1})^T(\varphi_{i-1} + \xi_i)) = \\
&= \mathrm{u}(-\varphi_{i-2}^T\varphi_{i-1} - \varphi_{i-2}^T\xi_i - \varphi_{i-1}^T\xi_{i-1} - \xi_{i-1}^T\xi_i) = \\
&= \mathrm{u}(-\xi_{i-1}^T\xi_i) + \mathrm{u}'(\theta_i) \times \left(-\varphi_{i-2}^T\varphi_{i-1} - \varphi_{i-2}^T\xi_i - \varphi_{i-1}^T\xi_{i-1}\right)
\end{aligned} \tag{2.114}$$

onde θ_i é um ponto entre $-y_{i-1}^T y_i$ e $-\xi_{i-1}^T\xi_i$. Temos também que a função u' é limitada e $\varphi(x_i) \to 0$ de onde, pelo corolário do Lema de Kronecker (por exemplo, [48]),

$$\sum_{i=1}^{t} \mathrm{u}'(\theta_i)\varphi_{i-2}^T\varphi_{i-1} = o(t) \tag{2.115}$$

$$\sum_{i=1}^{t} \mathrm{u}'(\theta_i)\varphi_{i-2}^T\xi_i = o(t) \tag{2.116}$$

$$\sum_{i=1}^{t} \mathrm{u}'(\theta_i)\varphi_{i-1}^T\xi_{i-1} = o(t)\,. \tag{2.117}$$

Assim temos

$$\begin{aligned}
s_t &= s_0 + s_1 + \sum_{i=1}^{t}(\mathrm{u}(-y_{i-1}^T y_i) - \mathrm{u}(-\xi_{i-1}^T\xi_i)) + \\
&\quad + \sum_{\text{pares}}^{t} \mathrm{u}(-\xi_{i-1}^T\xi_i) + \sum_{\text{ímpares}}^{t} \mathrm{u}(-\xi_{i-1}^T\xi_i) \\
&= s_0 + s_1 + \Delta U_t + P_t + I_t.
\end{aligned}$$

Por (2.115), (2.116) e (2.117)

$$\Delta U_t = \sum_{i=1}^{t}(\mathrm{u}(-y_{i-1}y_i) - \mathrm{u}(-\xi_{i-1}\xi_i)) = o(t) \text{ quase-certamente}.$$

Cada uma das somas P_t e I_t é composta por parcelas independentes de média E_0 e variância finita. Pela lei do logaritmo iterado

$$P_t + I_t = \mathrm{E}_0 t + \mathrm{O}(\sqrt{t \log\log t})\,.$$

Sendo $\lim_{t\to\infty} s_0/t = 0$ *quase-certamente*, idem para s_1, temos

$$s_t = s_0 + s_1 + \mathrm{E}_0 t + t\mathrm{o}_t + \mathrm{O}(\sqrt{t \log\log t}) = (\mathrm{E}_0 + \mathrm{o}_t)t,$$

quase-certamente. Então

$$\begin{aligned} s_t &= (\mathrm{E}_0 + \mathrm{o}_t)t = \mathrm{E}_0 t \left(\frac{1}{1 - \frac{\mathrm{o}_t}{\mathrm{E}_0 + \mathrm{o}_t}}\right) = \\ &= \mathrm{E}_0 t \left(\frac{1}{1 + \mathrm{o}_t}\right). \end{aligned}$$

□

Demonstração do Teorema 4 (Normalidade assimptótica) Fazemos $x^* = 0$. A convergência *quase-certa* de $x_t \to 0$ demonstrada na Secção anterior, a convergência *quase-certa* de $o_t \to 0$ sendo esta v.a. definida no Lema 19 e o Lema 9 na página 32 (Delyon e Juditsky [14]), conduzem a que exista uma sequência (a_t) de números não aleatórios positivos tal que

$$a_t \to 0 \quad \text{e} \quad |\mathrm{o}_t|/a_t \to 0, \quad |x_t|/a_t \to 0 \quad \textit{quase-certamente}. \tag{2.118}$$

Comentário 20 *Recordamos a explicação dada no caso unidimensional. A explicação para o acima dado é que se fizermos $\theta_t := |o_t| + |x_t|$ então $\theta_t \to 0$ quase-certamente. Então existe $a_t \to 0$, deterministicamente, tal que $\theta_t/a_t \to 0$ quase-certamente. Daqui segue que $|o_t|/a_t \to 0$ e $|x_t|/b_t \to 0$ quase-certamente.*

Definimos os tempos-de-paragem

$$\tau_R = \inf\{t : |\mathrm{o}_t| \geq R|a_t|\}, \quad \sigma_R = \inf\{t : |x_t| \geq R|a_t|\} \tag{2.119}$$

para $R > 0$ e

$$\nu = \min(\tau_R, \sigma_R)\,. \tag{2.120}$$

Do Lema 9 e de (2.118) concluímos que para qualquer $\epsilon > 0$ podemos escolher $R < \infty$ tal que

$$\mathrm{P}(\nu = \infty) \geq 1 - \epsilon. \tag{2.121}$$

Desta maneira, com uma probabilidade tão grande quanto se deseje, temos um majorante determinístico comum a $|o_t|$ e $|x_t|$ e que será usado na demonstração.

Consideramos o processo análogo ao algoritmo em (2.73) mas com passo determinístico $\gamma_t = 1/(\mathrm{E}_0 t)$ e aplicado à função $\varphi(x) = \alpha x$ (α é a derivada de φ em x^*),

$$z_t = z_{t-1} - \frac{1}{\mathrm{E}_0 t}(\alpha z_{t-1} + \xi_t), \quad z_0 = x_0. \tag{2.122}$$

As propriedades assimptóticas deste processo são conhecidas (ver, por exemplo, Nevel'son e Has'minskii [31]). Assim

$$\begin{aligned} z_t t^{1/2-\epsilon} \to 0, \textit{quase-certamente}, \text{ para todo } \epsilon > 0, \\ \mathrm{E}|z_t|^2 \leq K/t, \quad K > 0 \\ \sqrt{t} z_t \xrightarrow{\mathrm{d}} N(0, V). \end{aligned} \tag{2.123}$$

onde V é a matriz definida em (2.78).

Baseado no Lema 23, o Lema 21 demonstrará que, assimptoticamente, $\sqrt{t}x_t$ e $\sqrt{t}z_t$ têm a mesma distribuição limite, descrita em (2.123). □

Lema 20 *Seja A uma matriz definida positiva e simétrica, a, b, c e d vectores reais. Então*

$$\begin{aligned} (a+b+c+d)^T A(a+b+c+d) \leq\ & a^T Aa + \\ & +3(b^T Ab + c^T Ac + d^T Ad) + \\ & +a^T Ab + b^T Aa + \\ & +2a^T A(c+d) \end{aligned}$$

Prova. De

$$\begin{aligned} (a-b)^T A(a-b) &= a^T Aa + b^T Ab - a^T Ab - b^T Aa \geq 0 \Leftrightarrow \\ &\Leftrightarrow a^T Ab + b^T Aa \leq a^T Aa + b^T Ab \end{aligned}$$

temos

$$\begin{aligned} (a+b)^T A(a+b) &= a^T Aa + b^T Ab + a^T Ab + b^T Aa \\ &\leq a^T Aa + b^T Ab + a^T Aa + b^T Ab \\ &= 2(a^T Aa + b^T Ab) \end{aligned}$$

De modo semelhante

$$\begin{aligned}(a+b+c)^T A(a+b+c) &= a^T Aa + b^T Ab + c^T Ac + \\ &\quad (a^T Ab + b^T Aa) + (a^T Ac + c^T Aa) + \\ &\quad (b^T Ac + c^T Ab) \\ &\leq a^T Aa + b^T Ab + c^T Ac + \\ &\quad (a^T Aa + b^T Ab) + (a^T Aa + c^T Ac) + \\ &\quad (b^T Ab + c^T Ac) \\ &= 3(a^T Aa + b^T Ab + c^T Ac)\end{aligned}$$

Então, temos

$$\begin{aligned}(a+b+c+d)^T A(a+b+c+d) &= (a+(b+c+d))^T A(a+(b+c+d)) \\ &= a^T Aa + \\ &\quad a^T A(b+c+d) + \\ &\quad (b+c+d)^T Aa + \\ &\quad (b+c+d)^T A(b+c+d) \\ &\leq a^T Aa + \\ &\quad 3(b^T Ab + c^T Ac + d^T Ad) + \\ &\quad a^T Ab + b^T Aa + \\ &\quad 2a^T A(c+d)\end{aligned}$$

□

Lema 21 *Seja* $\Delta_t := x_t - z_t$. *Então* $\sqrt{t}\Delta_t \xrightarrow{pr} 0$.

Prova. Do Lema 19, $\gamma_t = \frac{1}{s_t} = \frac{1}{E_0 t}(1+o_t)$ onde o_t é uma v.a. independente de ξ_{t+1} que converge para 0 *quase-certamente*. Então escrevemos

$$x_{t+1} = x_t - \frac{1}{E_0 t}(1+o_t)(\varphi(x_t) + \xi_{t+1}) \tag{2.124}$$

e

$$\begin{aligned} x_{t+1} = x_t &- \frac{1}{E_0 t}\varphi(x_t) - \\ &- \frac{1}{E_0 t}\xi_{t+1} - \\ &- \frac{o_t}{E_0 t}\varphi(x_t) - \\ &- \frac{o_t}{E_0 t}\xi_{t+1} \end{aligned}$$

Consideramos, sem perda de generalidade, que $\varphi(0) = 0$ e da Condição B3.4 usamos a expansão em série de Taylor,

$$\varphi(x) = (\varphi(x) - \varphi'(0)x) + \varphi'(0)x$$

e então

$$\begin{aligned} x_{t+1} = x_t &- \frac{1}{E_0 t}\varphi'(0)x_t - \\ &- \frac{1}{E_0 t}\xi_{t+1} - \\ &- \frac{o_t}{E_0 t}\xi_{t+1} - \\ &- \frac{1}{E_0 t}\left(o_t\varphi(x_t) + \varphi(x_t) - \varphi'(0)x_t\right) . \end{aligned}$$

Definimos

$$v_t := o_t\frac{\varphi(x_t)}{|x_t|} + \frac{\varphi(x_t) - \varphi'(0)x_t}{|x_t|}$$

e para $t \leq \nu$ temos que $|x_t| \leq Ra_t$ e $|o_t| \leq Ra_t$

$$\begin{aligned} |v_t| &\leq Ra_t \sup_x \frac{|\varphi(x)|}{|x|} + \sup_{|x|\leq Ra_t} \frac{|\varphi(x_t) - \varphi'(0)x_t|}{|x_t|} \leq \\ &\leq Ra_t M + o(1) := c_t . \end{aligned} \tag{2.125}$$

Notamos que $c_t \to 0$ onde c_t é uma sequência positiva decrescente e então

$$\begin{aligned} x_{t+1} = x_t &- \frac{1}{E_0 t}\varphi'(0)x_t - \\ &- \frac{1}{E_0 t}\xi_{t+1} - \\ &- \frac{o_t}{E_0 t}\xi_{t+1} - \\ &- \frac{1}{E_0 t}v_t|x_t| . \end{aligned}$$

Consideramos o algoritmo para z_t

$$\begin{aligned} z_{t+1} &= z_t - \frac{1}{E_0 t}(\varphi'(0)z_t + \xi_{t+1}) = \\ &= z_t - \frac{1}{E_0 t}\varphi'(0)z_t - \frac{1}{E_0 t}\xi_{t+1} \end{aligned}$$

e

$$\begin{aligned} x_{t+1} &= x_t - \frac{1}{E_0 t}\varphi'(0)x_t - \frac{1}{E_0 t}\xi_{t+1} - \frac{o_t}{E_0 t}\xi_{t+1} - \frac{1}{E_0 t}v_t|x_t| \\ z_{t+1} &= z_t - \frac{1}{E_0 t}\varphi'(0)z_t - \frac{1}{E_0 t}\xi_{t+1} \end{aligned}$$

de onde

$$\Delta_{t+1} = \Delta_t - \frac{1}{E_0 t}\varphi'(0)\Delta_t - \frac{1}{E_0 t}v_t|x_t| - \frac{o_t}{E_0 t}\xi_{t+1}\,.$$

Desejamos mostrar que $\sqrt{t}\Delta_t = \sqrt{t}(x_t - z_t) \stackrel{\text{pr}}{\to} 0$. Definimos $V_t := \Delta_t^T A \Delta_t$ para uma matriz definida positiva A a ser especificada.

Desejamos mostrar também que $\mathrm{E}[tV_t\ \mathbb{I}(t < \nu)] \to 0$ e pelo Teorema 7, página 62, vamos verificar que $\sqrt{t}(x_t - z_t) \stackrel{\text{pr}}{\to} 0$. Assim,

$$\begin{aligned} V_{t+1} &= \Delta_{t+1}^T A \Delta_{t+1} \\ &= (\Delta_t - \frac{1}{E_0 t}\varphi'(0)\Delta_t - \frac{1}{E_0 t}v_t|x_t| - \frac{o_t}{E_0 t}\xi_{t+1})^T \cdot \\ &\quad A \cdot \\ &\quad (\Delta_t - \frac{1}{E_0 t}\varphi'(0)\Delta_t - \frac{1}{E_0 t}v_t|x_t| - \frac{o_t}{E_0 t}\xi_{t+1}) \end{aligned}$$

ou, após transposição,

$$\begin{aligned} V_{t+1} &= \Delta_{t+1}^T A \Delta_{t+1} \\ &= (\Delta_t^T - \frac{1}{E_0 t}\Delta_t^T\varphi'(0)^T - \frac{1}{E_0 t}v_t^T|x_t| - \frac{o_t}{E_0 t}\xi_{t+1}^T)^T \cdot \\ &\quad A \cdot \\ &\quad (\Delta_t - \frac{1}{E_0 t}\varphi'(0)\Delta_t - \frac{1}{E_0 t}v_t|x_t| - \frac{o_t}{E_0 t}\xi_{t+1}) \end{aligned}$$

Para estimar V_{t+1} usamos o Lema 20 e obtemos

$$V_{t+1} \le V_t + B_t + C_t + D_t$$

com B_t, C_t e D_t a ser especificado. Com, $\mathbb{I}(t+1 < \nu) \le \mathbb{I}(t < \nu)$, procuramos simplificar

$$\begin{aligned} \mathrm{E}[(t+1)V_{t+1}\ \mathbb{I}(t+1 < \nu)] \le\ & \mathrm{E}[(t+1)V_t\ \mathbb{I}(t < \nu)] \\ & +\mathrm{E}[(t+1)B_t\ \mathbb{I}(t < \nu)] \\ & +\mathrm{E}[(t+1)C_t\ \mathbb{I}(t < \nu)] \\ & +\mathrm{E}[(t+1)D_t\ \mathbb{I}(t < \nu)] \end{aligned}$$

obtendo uma estimativa de $tV_t\ \mathbb{I}(t < \nu)$ de modo a provar a convergência para zero em esperança matemática.

Vamos considerar tempos $t \leq \nu$ e então $|x_t| \leq Ra_t$ e $|o_t| \leq Ra_t$. Primeiro obtemos B_t

$$\begin{aligned} B_t &= \frac{3}{E_0^2 t^2}\left(\Delta_t^T \varphi'(0)^T A \varphi'(0)\Delta_t + |x_t|^2 v_t^T A v_t + o_t^2 \xi_{t+1}^T A \xi_{t+1}\right) \\ &\leq \frac{3}{E_0^2}\frac{1}{t^2}\left(K_1 \cdot V_t + |v_t|^2 \cdot |x_t|^2 \cdot |A| + o_t^2 |A| |\xi_{t+1}|^2\right) \\ &\leq \frac{3}{E_0^2}\frac{1}{t^2}\left(K_1 \cdot V_t + c_t^2 \cdot R^2 a_t^2 \cdot |A| + R^2 a_t^2 \cdot |\xi_{t+1}|^2 \cdot |A|\right) \\ &\leq \frac{3}{E_0^2}\frac{1}{t^2}\left(K_1 \cdot V_t + o(1) + o(1) \cdot |\xi_{t+1}|^2\right) \end{aligned}$$

onde K_1 é uma constante positiva tal que

$$\Delta_t^T \varphi'(0)^T A \varphi'(0)\Delta_t \leq K_1 \Delta_t^T A \Delta_t = K_1 V_t.$$

Simplificando

$$\begin{aligned} \mathrm{E}[(t+1)B_t\, \mathbb{I}(t < \nu)] &\leq \mathrm{E}[(t+1)B_t]\mathrm{P}(t < \nu) \\ &\leq \mathrm{E}[(t+1)B_t]\,. \end{aligned}$$

Primeiro

$$(t+1)B_t \leq \frac{3(t+1)}{E_0^2}\frac{1}{t^2}\left(K_1 \cdot V_t + o(1) + o(1) \cdot |\xi_{t+1}|^2\right)$$

e porque

- $\frac{3(t+1)}{E_0^2}\frac{1}{t^2} \leq \frac{K_3}{t}$, para alguma constante positiva K_3;
- $\frac{3(t+1)}{E_0^2}\frac{1}{t^2} o(1) = o(t^{-1})$;
- $\mathrm{E}[|\xi_{t+1}|^2] = tr(S_\xi)$;

temos

$$\mathrm{E}[(t+1)B_t] = \frac{K_3}{t} V_t + o(t^{-1})\,.$$

Desenvolvemos agora C_t,

$$\begin{aligned} C_t &= \Delta_t^T A \frac{-1}{E_0 t}\varphi'(0)\Delta_t + \frac{-1}{E_0 t}\Delta_t^T \varphi'(0) A \Delta_t = \\ &= \frac{-1}{t}\Delta_t^T (A\varphi'(0)/E_0 + \varphi'(0)^T/E_0 A)\Delta_t\,. \end{aligned}$$

De modo a estimar C_t de um modo útil determinamos uma matrix A que verifique $A\varphi'(0)/E_0 + \varphi'(0)^T/E_0 A = I + A$ e ainda $I + A \geq (1+\beta)A$ para uma constante real positiva β. Seguem os detalhes. O valor β segue de

$$\begin{aligned} I + A &\geq (1+\beta)A \Leftrightarrow \\ &\Leftrightarrow I \geq \beta A\,. \end{aligned}$$

Escrevemos, para $A = A^T$,

$$\begin{aligned} A\varphi'(0)/E_0 + \varphi'(0)^T/E_0 A &= I + A \Leftrightarrow \\ \varphi'(0)^T/E_0 A + A\varphi'(0)/E_0 &= I + A \Leftrightarrow \\ \varphi'(0)^T/E_0 A - \frac{A}{2} + A\varphi'(0)/E_0 - \frac{A}{2} &= I \Leftrightarrow \\ (\varphi'(0)^T/E_0 - \frac{I}{2})A + A(\varphi'(0)/E_0 - \frac{I}{2}) &= I \end{aligned}$$

e para usarmos o resultado de Lyapunov (Teorema 5) reescrevemos a última igualdade da seguinte maneira

$$(\frac{I}{2} - \varphi'(0)^T/E_0)A + A(\frac{I}{2} - \varphi'(0)/E_0) = -I$$

onde, da Condição B3.3, $\frac{I}{2} - \varphi'(0)/E_0$ é definida negativa. Então, a solução A existe e é positiva definida. Então,

$$\begin{aligned} C_t &= \frac{-1}{t}\Delta_t^T(A\varphi'(0)/E_0 + \varphi'(0)^T/E_0 A)\Delta_t \\ &= \frac{-1}{t}\Delta_t^T(A + I)\Delta_t \\ &\leq -(1+\beta)\frac{1}{t}V_t \end{aligned}$$

Estimamos o último termo D_t

$$D_t = \frac{-1}{E_0 t}(2\Delta_t^T A v_t \cdot |x_t| + 2\Delta_t^T A o_t \xi_{t+1})\,.$$

Recordamos que estamos a considerar $t < \nu$ e porque não é válido que $|\Delta_t| \leq V_t$ seguimos o seguinte

- $x_t = \Delta_t + z_t$ de onde $|x_t|^2 \leq |\Delta_t|^2 + |z_t|^2$;
- $2|\Delta_t|^2 \leq K_2 V_t$ (2 por conveniência) para uma certa constante positiva K_2.

Então,

$$\begin{aligned} 2\Delta_t^T A v_t \cdot |x_t| &\leq 2|\Delta_t| \cdot |x_t| \cdot |A| \cdot c_t \\ &\leq (|\Delta_t|^2 + |x_t|^2) \cdot |A| \cdot c_t \\ &\leq (2|\Delta_t|^2 + |z_t|^2) \cdot |A| \cdot c_t \\ &\leq (K_2 V_t + |z_t|^2) \cdot |A| \cdot c_t \end{aligned}$$

Regressando à estimação de D_t

$$\begin{aligned} D_t &\leq \frac{-1}{E_0 t}(2\Delta_t^T A v_t \cdot |x_t| + 2\Delta_t^T A o_t \xi_{t+1}) \leq \\ &\leq \frac{K_2}{E_0 t} \cdot |A| \cdot c_t \cdot V_t \\ &\quad + \frac{1}{E_0 t} \cdot |A| \cdot c_t \cdot |z_t|^2 \\ &\quad - \frac{2}{E_0 t} \Delta_t^T A o_t \xi_{t+1} \,. \end{aligned}$$

Tomando

- $\mathrm{E}[|z_t|^2] = K_4/t$, para alguma constante positiva K_4;

Então

$$\begin{aligned} \mathrm{E}[(t+1)D_t] &= \frac{K_2(t+1)}{E_0 t} \cdot |A| \cdot c_t \cdot V_t \\ &\quad + \frac{t+1}{E_0 t} \cdot |A| \cdot c_t \cdot \frac{K_4}{t} \\ &\leq o(1)V_t + o(t^{-1}) \end{aligned}$$

Agora juntando tudo, sempre considerando $t < \nu$,

$$\begin{aligned} (t+1)V_{t+1} &\leq (t+1)V_t + \frac{K_3}{t}V_t + \\ &\quad o(t^{-1}) - \frac{t+1}{t}(1+\beta)V_t + \\ &\quad o(1)V_t + o(t^{-1}) \\ &\leq V_t(t+1\frac{K_3}{t} - (1+\beta)\frac{t+1}{t} + o(1)) + o(t^{-1}) \\ &\leq t \cdot V_t(1 + \frac{1}{t} + \frac{K_3}{t^2} - (1+\beta)\frac{t+1}{t^2} + o(t^{-2})) + o(t^{-1}) \\ &\leq tV_t(1 - (1+\beta)\frac{1}{t} + o(t^{-1})) + o(t^{-1}) \\ &\leq tV_t(1 - (1+\beta+o(1))\frac{1}{t}) + o(t^{-1}) \\ &\leq tV_t(1 - (\beta/2)\frac{1}{t}) + o(t^{-1}) \,. \end{aligned}$$

Segue que,

$$\mathrm{E}[(t+1)V_{t+1}\, \mathbb{I}(t+1 < \nu)] \leq \mathrm{E}[tV_t\, \mathbb{I}(t < \nu)] + o(t^{-1})$$

e pelo Lema 20

$$\mathrm{E}[tV_t\, \mathbb{I}(t < \nu)] \to 0$$

então pelo Teorema 7

$$tV_t\, \mathbb{I}(t < \nu) \xrightarrow{\mathrm{pr}} 0$$

ou

$$\sqrt{t}(x_t - z_t)\, \mathbb{I}(t < \nu) \xrightarrow{\text{pr}} 0\,,$$

ou ainda, por definição da convergência em probabilidade,

$$\forall \eta > 0 \quad \mathrm{P}(|\sqrt{t}(x_t - z_t)\, \mathbb{I}(t < \nu)| < \eta) \to 1\,.$$

Os eventos seguintes estão relacionados por

$$\sqrt{t}(x_t - z_t) < \eta \Rightarrow \sqrt{t}(x_t - z_t)\, \mathbb{I}(t < \nu) < \eta$$

e por $P(\sqrt{t}(x_t - z_t) < \eta) \leq \mathrm{P}(\sqrt{t}(x_t - z_t)\, \mathbb{I}(t < \nu) < \eta)$ temos

$$\sqrt{t}(x_t - z_t) \xrightarrow{\text{pr}} 0\,.$$

□

2.7 Resultados padrão usados

Teorema 5 (A. M. Lyapunov, 1947 (citado em [26], Cap. 13.1)) *Seja $U, W \in \mathbb{C}^{n\times n}$ e seja W definida positiva.*

(a) Se U é estável então a equação

$$UA + AU^* = W$$

tem uma única solução A definida negativa.

(b) Se existir uma matriz definida negativa A satisfazendo a equação acima então A é estável.

Comentário 21 *Estável é quando os valores próprios são todos negativos. Quando os valores próprios são todos negativos então a matriz é definida negativa. O termo estável ocorre em virtude da equação diferencial matricial.*

Lema 22 (Markov Inequality (por exemplo, [48])) *Seja Z uma v.a. e $g : \mathbb{R} \to [0, \infty]$ uma função não decrescente. Então*

$$\mathrm{E}g(Z) \geq \mathrm{E}(g(Z); Z \geq c) \geq g(c)\mathrm{P}(Z \geq c)$$

Teorema 6 (Martingale convergence, [48], Cap. 12) *Seja M um martingale para o qual $M_n \in \mathcal{L}^2, \forall n$. Então M é limitado em $\mathcal{L}^2$ se e só se*

$$\sum \mathrm{E}[(M_k - M_{k-1})^2] < \infty$$

e quando isto ocorre obtém-se

$$M_n \to M_\infty \text{ quase-certamente e em } \mathcal{L}^2 .$$

Teorema 7 ([48], Cap. 13.7) *Seja (X_n) uma sequência em $\mathcal{L}^1$ e $X \in \mathcal{L}^1$. Então $X_n \to X$ in $\mathcal{L}^1$, ou equivalentemente $\mathrm{E}(|X_n - X|) \to 0$, se e só se, as condições seguintes estão satisfeitas:*

1. *$X_n \to X$ em probabilidade;*
2. *a sequência (X_n) é uniformemente integrável ($\forall \epsilon > 0 \exists K : \mathrm{E}[|X|; |X| > K] < \epsilon$).*

Lema 23 (p. ex. [11, p.359]) *Se $|X_t - Z_t| \xrightarrow{pr} 0$ e X_t converge em distribuição então Z_t converge em distribuição para o mesmo limite.*

Teorema 8 (Kolmogorov Law of Iterated Logarithm (p.ex. [48])) *Sejam $X_1, X_2, \ldots$ variáveis aleatórias independentes e identicamente distribuídas com média 0 e variância 1. Faz-se $S_n := X_1 + \cdots + X_n$. Então, quase-certamente,*

$$\limsup \frac{S_n}{\sqrt{2n \log\log n}} \to +1, \qquad \liminf \frac{S_n}{\sqrt{2n \log\log n}} \to -1 .$$

Lema 24 *Consideramos o algoritmo de passo constante ρ*

$$x_t = x_{t-1} - \rho\varphi(x_{t-1})$$

onde $\rho \leq \gamma(0)$ sendo $\gamma(0)$ uma constante definida na Condição A3 da Secção 2.1, e em que $x_{t-1} \in \mathbb{R}$, $\varphi(x_{t-1}) \in \mathbb{R}$ é uma função real de variável real com um único zero (assumimos $\varphi(0) = 0$).

Prova. Vamos desenvolver

$$|x_t| = |x_0| \frac{|x_1|}{|x_0|} \cdots \frac{|x_t|}{|x_{t-1}|}$$

e de onde

$$\frac{x_t}{x_{t-1}} = 1 - \rho \frac{\varphi(x_{t-1})}{x_{t-1}} .$$

Da Condição A3 temos $\sup \varphi(x)/x = M < \infty$ e $\rho < 2/M$, e ainda $\varphi(x)/x > 0$ para $x \neq 0$. Assim

$$1 > 1 - \rho\frac{\varphi(x_{t-1})}{x_{t-1}} > 1 - \rho M > -1$$

de onde verificamos que

$$\frac{|x_t|}{|x_{t-1}|} < 1$$

Definimos $m := \inf_{|x|<|x_0|} \frac{\varphi(x)}{x}$. Basta considerar este ínfimo sobre os valores $|x| < |x_0|$ pois pela equação acima a sequência $|x_t|$ é decrescente. Falta verificar que decresce para zero. Assim

$$1 > 1 - \rho m > 1 - \rho\frac{\varphi(x_{t-1})}{x_{t-1}} > 1 - \rho M > -1$$

e definido $r := \max\{|1 - \rho m|, |1 - \rho M|\}$ temos

$$\frac{|x_t|}{|x_{t-1}|} < r < 1$$

ou

$$\begin{aligned} |x_t| &= |x_0|\frac{|x_1|}{|x_0|} \cdots \frac{|x_t|}{|x_{t-1}|} \\ &\leq |x_0| r^t \to 0\,. \end{aligned}$$

□

Capítulo 3

Adaptação multiplicativa do passo

3.1 Introdução

Relembramos a forma padrão do algoritmo de aproximação estocástica para encontrar o zero duma função cujo cálculo é sujeito a um erro. O algoritmo consiste em calcular as aproximações sucessivas, $x_0,\ x_1,\ x_2, \ldots$, de acordo com a regra

$$x_t = x_{t-1} - \gamma_{t-1} y_t, \quad t = 1,\ 2, \ldots, \tag{3.1}$$

onde

$$y_t = \varphi(x_{t-1}) + \xi_t \tag{3.2}$$

é o valor de φ medido em x_{t-1}, ξ_t é o erro da medida; $\gamma_0,\ \gamma_1,\ \gamma_2, \ldots$ é a sequência de passos do algoritmo. Habitualmente é assumido que o tamanho dos passos sejam números positivos que satisfaçam as relações $\sum \gamma_t = \infty,\ \sum \gamma_t^2 < \infty$.

Relembramos ainda o trabalho teórico de Kesten [23], em que foi considerado o algoritmo usando (3.1) e (3.2), com a regra de adaptação do passo γ_t:

$$\gamma_t = \gamma(s_t), \qquad s_t = \begin{cases} s_{t-1} & \text{se} \quad y_{t-1} y_t > 0 \\ s_{t-1} + 1 & \text{se} \quad y_{t-1} y_t \leq 0, \end{cases} \tag{3.3}$$

onde $s_0 \in \mathbb{N};\ \gamma(0),\ \gamma(1),\ \gamma(2), \ldots$ é uma sequência decrescente de números positivos satisfazendo as relações $\sum \gamma(m) = \infty,\ \sum \gamma^2(m) < \infty$. No decorrer do algoritmo o passo não pode crescer; só pode manter-se constante ou decrescer. É suposto existir um único zero em φ e Kesten provou que x_t converge *quase-certamente* para o zero de φ.

Existem trabalhos heurísticos (em particular, nas redes neuronais artificiais), onde o passo é multiplicado por uma constante positiva inferior a 1, se alguma medida sobre os dados

indica que x_t está perto dum zero de φ, e por uma constante superior a 1, no casos restantes [1, 2, 43, 44]. Este tipo de regras garante uma taxa de convergência mais elevada mas, no entanto, a sequência de passos converge como numa progressão geométrica, e portanto $\sum \gamma_t < \infty$, o que significa que o limite de $\{x_t\}$ pode não ser necessariamente um ponto nulo de φ mas antes que a sequência fica retida algures o seu percurso para um dos zeros de φ. Porém, justifica-se o uso desta técnica se o resultado for um ponto bastante próximo a um dos zeros de φ.

O algoritmo proposto para adaptação do passo no algoritmo padrão usa este princípio. Adiciona-se às regras (3.1) e (3.2) a seguinte regra

$$\gamma_t = \begin{cases} \min\{u\,\gamma_{t-1},\, \bar{\mathrm{g}}\} & \text{se} \quad y_{t-1}y_t > 0, \\ d\,\gamma_{t-1} & \text{se} \quad y_{t-1}y_t \le 0, \end{cases} \qquad t = 2, 3, \ldots \tag{3.4}$$

Aqui $0 < d < 1 < u,\ 0 < \gamma_0, \gamma_1 \le \bar{\mathrm{g}},\ \bar{\mathrm{g}}$ é uma constante positiva. Notemos que as diferenças principais entre (3.4) e a regra de Kesten (3.3). Primeiro, de acordo com (3.4), γ_t pode ser decrementado ou incrementado. Segundo, no algoritmo de Kesten garante-se sempre que $\sum \gamma_t = \infty$. Por outro lado, no caso de convergência do algoritmo (3.1), (3.2), e (3.4), talvez se possa inferir que γ_t convirja como uma progressão geométrica (esta conjectura será justificada na Secção 3.3), e por conseguinte o limite do algoritmo poderá não ser um zero de φ.

Supomos que $\{\xi_t\}$ é uma sequência de v.a. i.i.d. com média zero, e ainda que $\mathrm{P}(\xi_t > 0) = \mathrm{P}(\xi_t < 0)$. Sob condições adicionais sobre φ, ξ_t, e $\bar{\mathrm{g}}$, em baixo, o processo definido por (3.1), (3.2) e (3.4) diverge *quase-certamente* se $ud > 1$, e converge se $ud < 1$, mais, o limite de $\{x_t\}$ pertence a $\mathcal{U}(\frac{\ln u}{-\ln d})$, onde $\mathcal{U}(\lambda),\ 0 < \lambda < 1$, é uma família monótona decrescente de conjuntos de números reais, e todo o conjunto $\mathcal{U}(\lambda)$ contém o conjunto Z de zeros de φ, e $\partial(\mathcal{U}(\lambda), \mathrm{Z}) \to 0$ quando $\lambda \to 1^-$.

Comentário 22 *Por definição* $\partial(A, B) := \sup_{x\in A} \inf_{y\in B} |x-y|$ *para quaisquer dois conjuntos de números reais* A *e* B*. Como exemplo, se consideramos* $\partial(\mathcal{U}(\lambda), \mathrm{Z})$ *então para um ponto de* $x \in \mathcal{U}(\lambda)$ *será escolhido um ponto dos zeros* Z *que fique mais próximo de* x*; considerando todos os pontos de* $x \in \mathcal{U}(\lambda)$ *destas distâncias menores a* Z*, escolhe-se a maior.*

Esta propriedade é consequência do Teorema principal, que será enunciado na Secção 3.2 e demonstrado na Secção 3.3. Assim, ao ajustar os parâmetros u e d (por exemplo, fixando u e fazendo $d \to (1/u)^-$, i.e., para valores à esquerda de $1/u$), pode-se atingir o desejado nível de precisão do algoritmo; maior precisão é obtida com menor taxa de convergência.

3.2 Enunciado do resultado principal

Consideramos o algoritmo dado por (3.1), (3.2) e (3.4). A regra (3.4) significa que em cada instante do algoritmo, o passo é multiplicado por u ou por d mas garantindo que o resultado da multiplicação por u seja inferior a $\bar{\mathrm{g}}$; caso contrário, é atribuído ao passo o valor $\bar{\mathrm{g}}$ por forma a que o valor máximo do passo seja $\bar{\mathrm{g}}$.

A regra (3.4) pode ser escrita na forma

$$\begin{aligned} \ln\tilde{\gamma}_t &= \ln\gamma_{t-1} + \ln u \cdot \mathbb{I}(y_{t-1}y_t > 0) + \ln d \cdot \mathbb{I}(y_{t-1}y_t \leq 0), \\ \ln\gamma_t &= \min\{\ln\tilde{\gamma}_t, \ln\bar{\mathrm{g}}\}. \end{aligned} \tag{3.5}$$

Consideramos que as seguintes condições são válidas:

A1 Seja $\mathcal{F}_t$, $t = 0, 1, 2, \ldots$ a σ-álgebra gerada por x_i, γ_i, e ξ_i, $0 \leq i \leq t$; então ξ_{t+1} não depende de $\mathcal{F}_t$.

A2 Os valores ξ_t são identicamente distribuídos, com média zero e variância positiva: $\mathrm{E}\xi_t = 0$, $\mathrm{Var}\xi_t =: S > 0$.

A3 (a) Existe $L > 0$ tal que para qualquer intervalo $I \subset [-L, L]$, $\mathrm{P}(\xi_1 \in I) > 0$;
(b) $\mathrm{P}(\xi_1 = 0) = 0$.

A4 $\varphi \in \mathrm{C}^1(\mathbb{R})$ e $\sup_x |\varphi'(x)| =: M < \infty$.

A5 $\bar{\mathrm{g}} < 2/M$.

A6 Existe $R > 0$ tal que

(a) $x\varphi(x) > 0$ quando $|x| \geq R$, e

(b) $\inf_{|x|\geq R} \varphi^2(x) > \dfrac{\bar{\mathrm{g}}MS}{2 - \bar{\mathrm{g}}M}$.

Comentário 23 *De A4 e A6 (a) segue que o conjunto* Z *não é vazio e está contido em* $(-R, R)$.

Comentário 24 *Fazemos notar que A4–A6 garantem a convergência da parte determinística do algoritmo (3.1), (3.2) e (3.4) (ou seja, quando* $\xi_t \equiv 0$*). Sob estas condições, qualquer algoritmo determinístico* $x_t = x_{t-1} - \gamma_{t-1}\varphi(x_{t-1})$ *converge, para qualquer sequência* $\{\gamma_t\}$ *que verifique* $\gamma_t \leq \bar{\mathrm{g}}$.

Comentário 25 *As próximas definições são necessárias em virtude das condições sobre a distribuição das perturbações ξ_t serem bastante gerais. Como se irá ver, no caso da distribuição destas ser simétrica, as próximas definições e análise ficariam bastante simplificadas.*

Introduzimos as funções:

$$k_+(z) := \lim_{\epsilon \to 0^+} \sup\{\mathrm{P}((\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0),\ |\varphi_1 - z| < \epsilon,\ |\varphi_2 - z| < \epsilon\}, \tag{3.6}$$

$$k_-(z) := \lim_{\epsilon \to 0^+} \inf\{\mathrm{P}((\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0),\ |\varphi_1 - z| < \epsilon,\ |\varphi_2 - z| < \epsilon\}; \tag{3.7}$$

e temos que $k_+(z) \geq 1/2,\ 0 \leq k_\pm(z) \leq 1,\ \lim_{z\to\infty} k_\pm(z) = 1$.

Comentário 26 *Temos que $k_+(z) \geq 1/2$ porque a distribuição do produto $\xi_1\xi_2$ tem esperança zero e é simétrica e como tal $\mathrm{P}(\xi_1\xi_2 > 0) = 1/2$. Assim o supremo será minorado por $1/2$.*

Definimos os conjuntos de números reais

$$V_\pm^{(a)} := \{x : k_\pm(\varphi(x)) < a\}, \quad V_\pm^{[a]} := \{x : k_\pm(\varphi(x)) \leq a\}; \tag{3.8}$$

onde $V_+^{(a)} \subset V_-^{(a)},\ V_\pm^{(a)} \subset V_\pm^{[a]}$ para cada a.

Comentário 27 *Em virtude de $k_-(z) \leq k_+(z) < a$ o conjunto das soluções $\{z : k_+(z) < a\}$ está contido em $\{z : k_-(z) < a\}$.*

Observemos que $V_+^{(a)}$ é aberto e que se $x \in V_+^{(a)}$, então existe $\epsilon > 0$ tal que

$$\sup\{\mathrm{P}((\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0),\ |\varphi_1 - \varphi(x)| < \epsilon,\ |\varphi_2 - \varphi(x)| < \epsilon\} =: c < a.$$

Assim, para x' suficientemente perto de x temos $|\varphi(x') - \varphi(x)| < \varepsilon/2$, e

$$\sup\{\mathrm{P}((\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0),\ |\varphi_1 - \varphi(x')| < \epsilon/2,\ |\varphi_2 - \varphi(x')| < \epsilon/2\} \leq c < a.$$

Isto implica que $k_+(\varphi(x')) < a$, logo $x' \in V_+^{(a)}$.

Denotamos também

$$k := \frac{\ln(1/d)}{\ln(u/d)}. \tag{3.9}$$

Definimos Z como o conjunto de zeros de φ, ou seja, $\mathrm{Z} := \{x : \varphi(x) = 0\}$. Suponhamos que $x \in V_+^{(k)},\ x_{t-2} \in (x - \epsilon, x + \epsilon) \subset V_+^{(k)}$, e $\gamma_{t-2} < \epsilon$, onde ϵ é um pequeno número positivo. Então, com probabilidade perto de 1, x_{t-1} também pertence a uma pequena (ou

eventualmente grande) vizinhança de x contida em $V_+^{(k)}$, e tomando em conta (3.6) e (3.8), obtemos

$$\begin{aligned}&\mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon) = \\ &= \mathrm{P}((\varphi(x_{t-2})+\xi_{t-1})(\varphi(x_{t-1})+\xi_t) > 0 \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon) < k.\end{aligned}$$

Então, usando (3.5) e (3.9), obtemos

$$\begin{aligned}&\mathrm{E}[\ln\gamma_t - \ln\gamma_{t-1} \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon] \leq \\ &\ln u \cdot \mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon) + \ln d \cdot \mathrm{P}(y_{t-1}y_t \leq 0 \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon) \\ &< \ln u \cdot k + \ln d \cdot (1-k) = 0.\end{aligned}$$

Num certo sentido, o conjunto $V_+^{(k)}$ pode ser visto como o *domínio de decremento do passo*: se vários valores consecutivos de x_t pertencem a $V_+^{(k)}$ e estão suficientemente perto entre si, e se o primeiro termo da sequência dos correspondentes passos γ_t é suficientemente pequeno, então a sequência dos seus valores esperados $\mathrm{E}\gamma_t$ decresce.

Comentário 28 *Da análise anterior foi concluido que "$V_+^{(k)}$ pode ser visto como o domínio de decremento do passo" mas aproveitamos para notar que o enunciado do Teorema estabelece a "convergência quase-certa para um ponto de $V_-^{[k]}$" ou seja, um conjunto que contem $V_+^{(k)}$ e não especificando se efectivamente a convergência só ocorre para pontos de $V_+^{(k)}$.*

Agora, supomos que $x \in \mathbb{R} \setminus V_-^{[k]}$, $x_{t-2} \in (x-\epsilon, x+\epsilon) \subset \mathbb{R} \setminus V_-^{[k]}$, e que $\gamma_{t-2} < \epsilon$. Analogamente, para ϵ suficientemente pequeno, temos que

$$\mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon) > k,$$

e depois, usando novamente (3.5) e (3.9) e tomando em conta que para $\epsilon < \bar{\gamma}/u^2$, $\tilde{\gamma}_t = \gamma_t$, obtemos

$$\begin{aligned}&\mathrm{E}[\ln\gamma_t - \ln\gamma_{t-1} \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon] = \\ &\ln u \cdot \mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon) + \ln d \cdot \mathrm{P}(y_{t-1}y_t \leq 0 \,\Big|\, |x_{t-2}-x|<\epsilon, \gamma_{t-2}<\epsilon) \\ &> \ln u \cdot k + \ln d \cdot (1-k) = 0.\end{aligned}$$

Assim, o conjunto $\mathbb{R} \setminus V_-^{[k]}$ pode ser visto como o *domínio de incremento do passo*: se vários valores consecutivos de x_t pertencem a $\mathbb{R} \setminus V_-^{[k]}$ e estão suficientemente perto entre si, e se os correspondentes e iniciais valores de γ_t são suficientemente pequenos, então a sequência dos valores esperados $\mathrm{E}\gamma_t$ cresce.

Notemos que se $k > k_+(0)$ então, por (3.8), $\mathrm{Z} \subset V_+^{(k)}$, ou seja, todos os zeros de φ pertencem à região de decremento do passo. Por outro lado, se $k < \inf_z k_-(z)$ então $V_-^{[k]} = \emptyset$, o que significa que a região de incremento do passo coincide com $\mathbb{R}$.

Parece provável que no primeiro caso o algoritmo possa convergir, e no segundo caso, não possa. Esta conjectura é confirmada pelo Teorema seguinte, que é o resultado principal deste Capítulo.

Teorema 9 (Plakhov e Cruz [34]) *Considerem-se as Condições A1–A6 satisfeitas; considere-se o processo $\{x_t,\ \gamma_t\}$ definido por (3.1), (3.2), (3.4). Relembramos que $k = \frac{\ln(1/d)}{\ln(u/d)}$. Então*
(a) Se $k > k_+(0)$ então $\{x_t\}$ converge quase-certamente para um ponto de $V_-^{[k]} := \{x : k_-(\varphi(x)) \le k\}$.
(b) Se $k < \inf_z k_-(z)$ então $\{x_t\}$ diverge quase-certamente.

Suponhamos que $\mathrm{P}(\xi_1 = x) = 0$ para cada x real e que $\mathrm{P}(\xi_1 > 0) = \mathrm{P}(\xi_1 < 0)$. Então a função $k(\cdot) := k_+(\cdot)$ coincide com $k_-(\cdot)$, é contínua, e é dada por

$$k(z) = \mathrm{P}((z + \xi_1)(z + \xi_2) > 0);$$

$z = 0$ é o único mínimo de $k(\cdot)$, e $k(0) = \inf_z k(z) = 1/2$. Depois de algumas contas, podemos escrever a hipótese do Teorema nas formas (a) $ud < 1$, e (b) $ud > 1$. Denotamos $\mathcal{U}(\lambda) := V^{[\frac{1}{1+\lambda}]} = \{x : k(\varphi(x)) \le \frac{1}{1+\lambda}\}$; $\mathcal{U}(\lambda)$, $0 < \lambda < 1$ é uma família decrescente de conjuntos contendo Z e que tende para Z quando $\lambda \to 1^-$.

Comentário 29 *Os conjuntos $V_\pm^{[}(k)]$ são monótonos com k. Da definição de k segue que $\frac{\ln u}{-\ln d} = \frac{1}{k} - 1$ e se k desce para $(1/2)^+$ então $\frac{\ln u}{-\ln d}$ sobe para 1^-. Assim, a subida monótona de $\frac{\ln u}{-\ln d}$ para 1^- causa uma diminuição monótona de k e consequentemente na amplitude dos conjuntos $V_\pm^{[}(k)]$.*

Assim, chegamos ao Corolário

Corolário 1 *Seja, em complemento das Condições A1–A6, $\mathrm{P}(\xi_1 = x) = 0$ para cada $x \in \mathbb{R}$, e $\mathrm{P}(\xi_1 > 0) = \mathrm{P}(\xi_1 < 0) = 1/2$. Consideramos o processo definido por (3.1), (3.2), (3.4). Então, existe uma família de conjuntos monótona decrescente $\mathcal{U}(\lambda)$, $0 < \lambda < 1$, tal que $\mathcal{U}(\lambda) \supset \mathrm{Z}$, $\partial(\mathcal{U}(\lambda), \mathrm{Z}) \to 0$ quando $\lambda \to 1^-$, e*
(a) se $ud < 1$ então $\{x_t\}$ converge quase-certamente para um ponto de $\mathcal{U}(\frac{\ln u}{-\ln d})$;
(b) se $ud > 1$ então $\{x_t\}$ diverge quase-certamente.

Comentário 30 *O Teorema não dá qualquer informação sobre o comportamento do algoritmo se u, d satisfaz as desigualdades*

$$\inf_z k_-(z) \leq \frac{\ln(1/d)}{\ln(u/d)} \leq k_+(0).$$

Em particular, sob as hipóteses do Corolário, o caso $ud = 1$ permanece por explorar.

3.3 Demonstração do Teorema

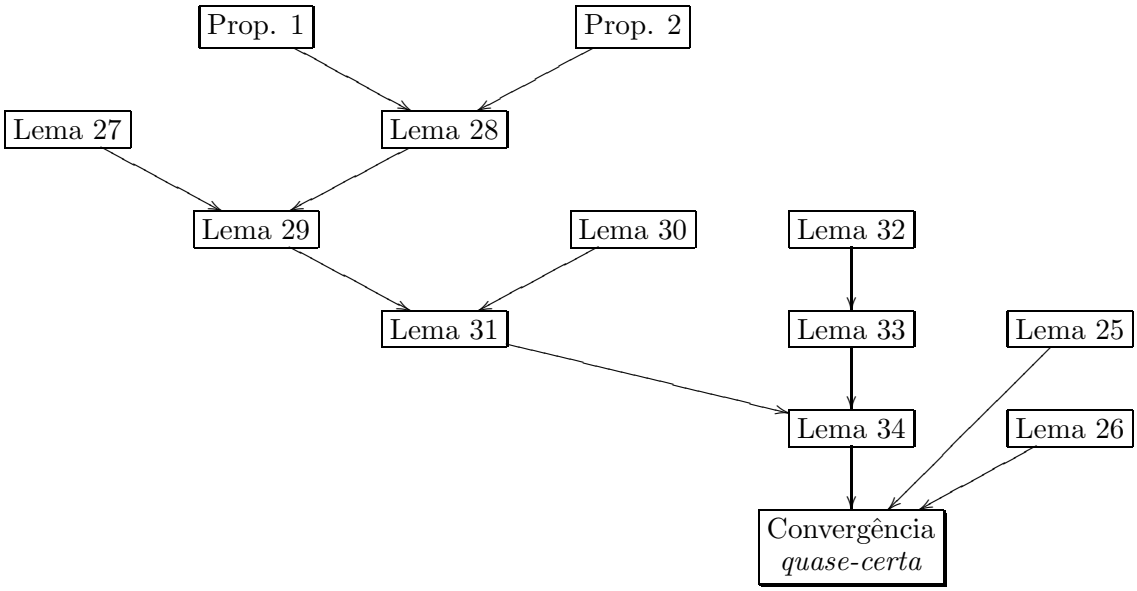


Figura 3.1: Lemas para a demonstração da convergência *quase-certa*.

A demonstração é constituída por 12 lemas auxiliares e baseada nestes segue a demonstração do Teorema. A Figura 3.1 ilustra as dependências entre Lemas e com base na mesma fazemos um resumo da demonstração.

Antes, relembramos que:

- A condição $k > k_+(0)$ compara a quantidade $k := k(u, d)$ que caracteriza o algoritmo e o valor $k_+(0)$ que depende apenas da distribuição das perturbações ξ_t. Esta condição, no caso das perturbações serem caracterizadas por uma distribuição com mediana em zero, condensa-se em $ud < 1$.

- No contexto de um único zero, se $\sum_t \gamma_t = \infty$ a sequência x_t pode alcançar o zero de φ e, intuitivamente, se $\sum_t \gamma_t < \infty$ a sucessão pode 'congelar' antes de alcançar o ponto desejado. Na demonstração são considerados os eventos $\{\sum_t \gamma_t = \infty\}$ e $\{\sum_t \gamma_t < \infty\}$.

A demonstração tem os seguintes passos principais:

- Se $ud < 1$ então $\sum_t \gamma_t < \infty$ (Lema 30).
- Se $\sum_t \gamma_t < \infty$ então $x_t \to x$ (Lema 21).
- Se x_t converge então $x \in \mathcal{U}(\frac{\ln u}{-\ln d}) \supset Z$ (Lema 22); relembramos que $\mathcal{U}(\frac{\ln u}{-\ln d})$ é um conjunto composto de vizinhanças dos zeros de φ.
- Se $ud > 1$ então $\mathcal{U}(\frac{\ln u}{-\ln d}) = \emptyset$, e x_t não pode convergir.
- Para eventos $\{\sum_t \gamma_t = \infty\}$ demonstra-se que para cada $\epsilon > 0$ e $g > 0$, *quase-certamente* existe t tal que $|x_t| < \epsilon$ e $\gamma_t < g$ (Lema 27). Este etapa da demonstração é usada no Lema 30, acima.

Fazemos agora uma exposição mais detalhada das etapas da demonstração começando pelo Lema 21.

As etapas da demonstração do Lema 21 enunciado acima são: mostrar que a sequência x_t é limitada para eventos $\{\sum_t \gamma_t < \infty\}$, e mostrar que sendo x_t uma sequência limitada, então converge. Para mostrar este facto, observamos

$$x_t = x_0 - \sum_{i=1}^{t} \gamma_{t-1}(\varphi(x_{t-1}) + \xi_t),$$

se x_t é limitado então $\varphi(x_t)$ também é limitado e por $\sum_t \gamma_t < \infty$ a parcela com φ converge; sobre a parcela com ξ_t usamos uma propriedade dos *martingales*: se $\sum_t E(z_t^2) < \infty$ então z_t converge *quase-certamente*, onde $z_t := \gamma_{t-1}\xi_t$, neste caso.

A demonstração do Lema 22 inicia-se supondo que uma sequência permanece numa vizinhança contida no complementar de $\mathcal{U}(\frac{\ln u}{-\ln d})$. Para tal sequência, verifica-se que ocorre $\gamma_{t-1} y_t > C > 0$ um número infinito de vezes, ou seja, é impossível a convergência nalgum ponto da vizinhança inicialmente escolhida.

Os próximos Lemas 23, 24, 25, 26 e 27 baseiam-se em eventos $\{\sum_t \gamma_t = \infty\}$ e as demonstrações são bastante próximas aos Lemas 1, 2, 3, 4, e 5 do caso algoritmo de Kesten Generalizado. Mas salientamos alguns aspectos que marcam diferenças importantes.

O Lema 24 mostra que com probabilidade positiva x_t alcança $\mathcal{O}$, se o processo partir de $|x_0| < R$. A demonstração do Lema 24 é análoga ao Lema 2 mas está dividida em 2 proposições: a Proposição 1 mostra que se o passo se mantém elevado então uma sequência, em que a amplitude das perturbações é controlada, alcança qualquer vizinhança $\mathcal{O}$ dos zeros de

φ; a Proposição 2 mostra que se duas iterações x_{t-1} e x_t estão fora de $\mathcal{O}$ então o passo aumenta $\gamma_{t+1} \geq \gamma_t$. Com base em ambas as proposições, o Lema 24 mostra que com probabilidade positiva x_t alcança $\mathcal{O}$.

O Lema 25 é análogo ao Lema 3 e a demonstração segue exactamente o mesmo raciocínio levando a que se conclua com probabilidade 1, que existe t tal que x_t está numa qualquer vizinhança $\mathcal{O}$ dos zeros de φ.

O Lema 26 mostra que se x_0 está numa certa vizinhança $\mathcal{O}$ então com probabilidade positiva, x_t estará numa vizinhança desejada $\mathcal{O}_1$ e com um passo tão pequeno quanto se deseja. A demonstração baseia-se no número de passos requerido para γ_t ir do passo máximo $\bar{g}$ ao passo desejado $\gamma_t = w$, e também, na distância máxima entre pontos de $\mathcal{O}$ e $\mathcal{O}_1$.

O Lema 27 é análogo ao Lema 5 e a demonstração segue exactamente o mesmo raciocínio levando a que se conclua com probabilidade 1, que existe t tal que x_t está numa qualquer vizinhança $\mathcal{O}$ dos zeros de φ e com um passo tão pequeno quando se deseja.

O Lema 28 estabelece que $\ln \gamma_t$ decresce pelo menos linearmente com probabilidade próxima de 1, quando x_0 parte duma região $\{x : |\varphi(x)| < \epsilon_0\}$. Na demonstração é construída uma sequência $\{\sigma_t\}$ tal que $\gamma_t < \sigma_t$, e em que σ_t se baseia apenas na sequência das perturbações ξ_t. A construção da sequência $\{\sigma\}$ satisfaz que se $\xi > 0$ então $\varphi(x) + \xi > 0$ quando $|\varphi| < \epsilon_0$. Deste modelo segue que se $y_{t-1}y_t < 0$ então $\xi_{t-1}\xi_t < 0$ e $|\xi_{t-1}| > \epsilon_0$ e $|\xi_t| > \epsilon_0$, ou seja,

$$\mathbb{I}(y_{t-1}y_t \geq 0) \leq \mathbb{I}(\xi_{t-1}\xi_t \geq 0 \text{ ou } |\xi_{t-1}| < \epsilon_0 \text{ ou } |\xi_t| > \epsilon_0)$$

causando $\gamma_t < \sigma_t$.

Os próximos Lemas 29 e 30 são estabelecidos quando os parâmetros u e d são tais que $ud < 1$ no caso das condições do Corolário 1, ou $k > k_+(0)$ no caso das condições mais gerais e enunciado do Teorema.

O Lema 29, o análogo do Lema 7 no caso de Kesten Generalizado, estabelece que quando $k > k_+(0)$ ($ud < 1$ no Corolário) há probabilidade positiva duma sequência x_t permanecer numa vizinhança dos zeros de φ com um passo que decresce pelo menos linearmente em logaritmo.

O Lema 30 estabelece que se $k > k_+(0)$ ($ud < 1$ no Corolário) então verifica-se com probabilidade 1 que $\sum_t \gamma_t < \infty$. O Lema 27 constata que

$$P(\text{se } \sum_t \gamma_t = \infty \text{ então } x_t \in \mathcal{O} \text{ e } \gamma_t < w) = 1,$$

para cada $\mathcal{O}$ e w dados. Esta proposição é equivalente a

$$P(\sum_t \gamma_t = \infty \text{ e } (x_t \notin \mathcal{O} \text{ ou } \gamma_t > w)) = 0$$

usada na demonstração do Lema 30.

Assumimos que todas as relações entre variáveis aleatórias são verdade *quase-certamente*.

Lema 25 *Se $\sum_t \gamma_t < \infty$ então a sequência $\{x_t\}$ converge.*

Prova. Notemos que sem perda de generalidade podemos assumir que x_0 é limitado. Repare-se que, substituindo x_0 por $\tilde{x}_0 = x_0 \cdot \mathbb{I}(|x_0| < X)$ muda o processo apenas com probabilidade $\mathrm{P}(|x_0| > X)$. Escolhendo X suficientemente grande, podemos tornar esta arbitrariamente pequena.

Seja $C > 0$; definimos o tempo-de-paragem $\tau_C = \inf\{t : \sum_{i=0}^{t} \gamma_i > C\}$ e introduzimos um novo processo x_t^C, γ_t^C definido por

$$\begin{aligned} x_t^C &= x_t, \quad \gamma_t^C = \gamma_t \text{ quando } t < \tau_c, \text{ e} \\ x_t^C &= x_{\tau_C}, \quad \gamma_t^C = 0 \text{ quando } t \geq \tau_c. \end{aligned}$$

Primeiro, vamos provar que a sequência $\{x_t^C\}$ é limitada. Designamos $M_R := \sup_{|x| \geq R} \frac{\varphi(x)}{x}$; de A4 segue que $M_R < \infty$. Temos que

$$|x_t^C| \leq |x_{t-1}^C - \gamma_{t-1}^C \varphi(x_{t-1}^C)| + \gamma_{t-1}^C |\xi_t|. \tag{3.10}$$

Usando $\gamma_{t-1}^C \leq C$ e $|\varphi(x_{t-1}^C)| \leq |\varphi(0)| + M|x_{t-1}^C|$, obtemos

$$|x_t^C| \leq |x_{t-1}^C|(1 + CM) + \gamma_{t-1}^C(|\varphi(0)| + |\xi_t|). \tag{3.11}$$

Se $\gamma_{t-1}^C \leq 2/M_R$, podemos obter uma estimativa ainda mais precisa para x_t^C. Distinguimos entre dois casos: (i) $|x_{t-1}^C| \leq R$ e (ii) $|x_{t-1}^C| > R$.

No caso (i), designamos $\bar{b} := \sup_{|x| \leq R} |\varphi(x)|$, e temos

$$|x_{t-1}^C - \gamma_{t-1}^C \varphi(x_{t-1}^C)| \leq |x_{t-1}^C| + \gamma_{t-1}^C \bar{b}. \tag{3.12}$$

No caso (ii) temos

$$0 \leq \gamma_{t-1}^C \frac{\varphi(x_{t-1}^C)}{x_{t-1}^C} \leq \frac{2}{M_R} M_R = 2,$$

portanto

$$|x_{t-1}^C - \gamma_{t-1}^C \varphi(x_{t-1}^C)| \leq |x_{t-1}^C|. \tag{3.13}$$

Assim, em ambos os casos (i) e (ii), de (3.10), (3.12), e (3.13) obtemos

$$|x_t^C| \leq |x_{t-1}^C| + \gamma_{t-1}^C(\bar{b} + |\xi_t|). \tag{3.14}$$

O número total de instantes t tal que $\gamma_{t-1}^C \leq 2/M_R$ é inferior a $CM_R/2$; por conseguinte, usando (3.11) e (3.14), podemos concluir que

$$|x_t^C| \leq \left(|x_0| + \sum_1^t \gamma_{i-1}^C(\bar{b} + |\varphi(0)| + |\xi_i|) \right) \cdot (1 + CM)^{CM_R/2}. \tag{3.15}$$

Denotamos $c_0 := \bar{b} + |\varphi(0)| + \mathrm{E}|\xi_1|$ e $\zeta_t := |\xi_t| - \mathrm{E}|\xi_t|$; usando que $\sum_1^\infty \gamma_{i-1}^C \leq C$ obtemos

$$|x_t^C| \leq \left(|x_0| + C\, c_0 + \sum_1^t \gamma_{i-1}^C \zeta_i \right) \cdot (1 + CM)^{CM_R/2}. \tag{3.16}$$

Usando que $\sum_1^\infty \mathrm{E}(\gamma_{t-1}^C \zeta_t)^2 = \mathrm{E}\zeta_1^2 \cdot \sum_1^\infty \mathrm{E}(\gamma_{t-1}^C)^2 < \infty$, obtemos que o *martingale* $\sum_1^t \gamma_{i-1}^C \zeta_i$ é limitado; o valor x_0 é também limitado, então, por (3.16), concluímos que a sequência $\{x_t^C\}$ é limitada.

Agora, vamos mostrar que $\{x_t^C\}$ converge. Da definição de x_t^C e γ_t^C segue que

$$x_t^C = x_0 - \sum_1^t \gamma_{i-1}^C \varphi(x_{i-1}^C) - \sum_1^t \gamma_{i-1}^C \xi_i.$$

Usando que a sequência $\{\varphi(x_{i-1}^C)\}$ é limitada e que $\sum_1^\infty \gamma_{i-1}^C \leq C$, obtemos que a série $\sum_1^\infty \gamma_{i-1}^C \varphi(x_{i-1}^C)$ converge. Mais, temos

$$\sum_1^\infty \mathrm{E}(\gamma_{t-1}^C \xi_t)^2 = S \cdot \sum_1^\infty \mathrm{E}(\gamma_{t-1}^C)^2 < \infty,$$

e portanto o *martingale* $\sum_1^t \gamma_{i-1}^C \xi_i$ converge. Isto implica que $\{x_t^C\}$ também converge.

Definimos o evento $A_C = \{\sum_t \gamma_t \leq C\}$ e $A_\infty = \{\sum_t \gamma_t < \infty\}$. Temos $A_\infty = \cup_C A_C$. Se $\sum_t \gamma_t \leq C$ então $x_t^C = x_t$ para cada t; isto significa que $\mathbb{I}(A_C) \cdot (x_t^C - x_t) = 0$ para cada t e C. A sequência $\{\, \mathbb{I}(A_C) x_t^C\}$ converge, portanto a sequência $\{\, \mathbb{I}(A_C) x_t\}$ também converge, e por passagem ao limite $C \to \infty$ obtemos que $\{\, \mathbb{I}(A_\infty) x_t\}$ converge. Isto significa exactamente que se $\sum_t \gamma_t < \infty$ então $\{x_t\}$ converge.

□

Lema 26 *Se $x_t \to x$ então $x \in V_-^{[k]}$.*

Prova. Usando A3 (a), é fácil verificar que existe $\delta_0 > 0$ tal que $\mathrm{P}(\xi_1 \notin [x-L/2,\, x+L/2]) > \delta_0$, para qualquer escolha de $x \in \mathbb{R}$.

De seguida, para cada $x \notin V_-^{([k])}$ existe $w(x) > 0$ e $0 < \epsilon(x) < L/4$ tal que o seguinte é válido: para quaisquer duas variáveis ϕ_1 e ϕ_2 satisfazendo $|\phi_l - \varphi(x)| \le \epsilon,\ l = 1, 2$ temos

$$\mathrm{P}((\phi_1 + \xi_1)(\phi_2 + \xi_2) > 0) > \frac{\ln(1/d) + w(x)}{\ln u + \ln(1/d)}.$$

Escolhe-se um conjunto de intervalos contável $U_i = (\varphi(x_i) - \epsilon(x_i),\ \varphi(x_i) + \epsilon(x_i))$ que cubra o conjunto $\varphi(\mathbb{R} \setminus V_-^{[k]})$, e denotamos $w_i := w(x_i)$. Fixamos i e $s \in \{0,\, 1,\, 2, \ldots\}$, e definimos o processo auxiliar $x_t^{(is)}$, $\gamma_t^{(is)}$ pelas fórmulas:

se $t < s$ então $x_t^{(is)} = x_t$, e se $t \ge s$ então

$$x_t^{(is)} = \begin{cases} x_{t-1}^{(is)} - \gamma_{t-1}^{(is)}\, y_t^{(is)} & \text{se } \ \varphi(x_{t-1}^{(is)} - \gamma_{t-1}^{(is)}\, y_t^{(is)}) \in U_i, \\ x_i & \text{outros casos;} \end{cases} \tag{3.17}$$

$$y_t^{(is)} = \varphi(x_{t-1}^{(is)}) + \xi_t, \tag{3.18}$$

$$\gamma_t^{(is)} = \begin{cases} \min\{u\gamma_{t-1}^{(is)},\, \bar{\mathrm{g}}\} & \text{se } \ y_{t-1}^{(is)}\, y_t^{(is)} > 0, \\ d\gamma_{t-1}^{(is)} & \text{se } \ y_{t-1}^{(is)}\, y_t^{(is)} \le 0. \end{cases} \tag{3.19}$$

Assim, quando $t \ge s,\ \varphi(x_t^{(is)})$ é forçado a estar contido em U_i.

Para $t \ge s + 2$, usando que $y_{t-1}^{(is)} = \varphi(x_{t-2}^{(is)}) + \xi_{t-1},\ y_t^{(is)} = \varphi(x_{t-1}^{(is)}) + \xi_t,\ \varphi(x_{t-2}^{(is)}) \in U_i$, obtemos que

$$\mathrm{P}(y_{t-1}^{(is)}\, y_t^{(is)} > 0) > \frac{\ln(1/d) + w_i}{\ln u + \ln(1/d)}$$

e

$$\mathrm{P}(y_{t-1}^{(is)}\, y_t^{(is)} \le 0) < \frac{\ln u - w_i}{\ln u + \ln(1/d)},$$

por isso

$$\begin{aligned} &\mathrm{E}[\ln u \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} > 0)\ +\ \ln d \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} \le 0)] > \\ &> \ln u \cdot \frac{\ln(1/d) + w_i}{\ln u + \ln(1/d)}\ +\ \ln d \cdot \frac{\ln u - w_i}{\ln u + \ln(1/d)} = w_i. \end{aligned}$$

Consideramos as variáveis $\phi_1 = f_1(\xi_1, \xi_2)$ e $\phi_2 = f_2(\xi_1, \xi_2)$ que fornecem a solução para o problema (determinístico) de minimização:

$$(\phi_1 + \xi_1)(\phi_2 + \xi_2) \to \min,$$

sujeito a

$$\begin{aligned} |\phi_1 - \varphi(x_i)| &\le \epsilon(x_i) \\ |\phi_2 - \varphi(x_i)| &\le \epsilon(x_i), \end{aligned}$$

e denotamos que $Y^1_{t-1} = f_1(\xi_{t-1}, \xi_t) + \xi_{t-1}$, $Y^2_t = f_2(\xi_{t-1}, \xi_t) + \xi_t$, $\eta_t = \ln u \cdot \mathbb{I}(Y^1_{t-1} Y^2_{t-1} > 0) + \ln d \cdot \mathbb{I}(Y^1_{t-1} Y^2_{t-1} \leq 0)$. Temos que

(i) $\eta_t \leq \ln u \cdot \mathbb{I}(y^{(is)}_{t-1}\, y^{(is)}_t > 0) + \ln d \cdot \mathbb{I}(y^{(is)}_{t-1}\, y^{(is)}_t \leq 0)$;

(ii) η_t são identicamente distribuídas, e $\mathrm{E}\eta_t \geq w_i$;

(iii) o conjunto de variáveis aleatórias $\{\eta_t,\ t \text{ pares},\ t \geq s+2\}$ tal como o conjunto $\{\eta_t,\ t \text{ ímpares},\ t \geq s+2\}$, são mutuamente independentes.

De (ii)–(iii) segue que *quase-certamente* $\sum_t \eta_t = +\infty$, e de (i) segue que

$$\sum_t [\ln u \cdot \mathbb{I}(y^{(is)}_{t-1}\, y^{(is)}_t > 0) + \ln d \cdot \mathbb{I}(y^{(is)}_{t-1}\, y^{(is)}_t \leq 0)] = +\infty,$$

então, em virtude de (3.19), $\gamma^{(is)}$ não converge para zero.

Assim, existe uma variável aleatória $\chi > 0$ tal que, para infinitos valores de t, $\gamma^{(is)}_t \geq \chi$.

Definimos a sequência de tempos-de-paragem $\tau_0, \tau_1, \tau_2, \ldots$ indutivamente, fazendo $\tau_0 = 0$ e $\tau_j = \inf\{t > \tau_{j-1} : \gamma^{(is)}_t \geq \chi\}$ para $j \geq 1$. Os eventos $B_j = \{|\xi_{\tau_j+1} + \varphi(x_i)| > L/2\}$ ocorrem com probabilidade superior a δ_0 (relembramos o comentário no início da demonstração), e todo o evento B_j, $j \geq 2$ não depende do conjunto de eventos $\{B_1, \ldots, B_{j-1}\}$. Assim, para infinitos valores de j, B_j ocorre, isto é, $|\xi_{\tau_j+1} + \varphi(x_i)| > L/2$, e então, considerando $|y_{\tau_j+1}| \geq |\xi_{\tau_j+1} + \varphi(x_i)| - |\varphi(x_{\tau_j}) - \varphi(x_i)|$ e $|\varphi(x_{\tau_j}) - \varphi(x_i)| < \epsilon(x_i) < L/4$, para estes valores de j temos $|y_{\tau_j+1}| \geq L/4$. Assim, concluímos que

$$\text{para uma quantidade infinita de valores de } j, \quad |\gamma_{\tau_j} y_{\tau_j+1}| \geq \chi\, L/4. \tag{3.20}$$

Supomos que x_t converge para um ponto de $\mathbb{R} \setminus V^{[k]}_-$, então para alguns i e s temos $x_t \in U_i$ quando $t \geq s$, e assim, o processo $x^{(is)}_t$, $\gamma^{(is)}_t$ coincide com x_t, γ_t, e de onde concluímos que $\gamma_t\, y_{t+1} \to 0$ quando $t \to \infty$. A última relação contradiz (3.20), e assim sendo o Lema 26 está demonstrado.

□

Lema 27 *Seja $\sum_t \gamma_t = \infty$. Então, para cada conjunto aberto $\mathcal{O}$ contendo* Z *existe uma constante positiva $g = g(\mathcal{O})$ tal que ocorre: (i) para alguns t, $x_t \in \mathcal{O}$, ou (ii) para alguns t, $|x_t| < R$ e $\gamma_t > g$.*

Prova. Designamos por f a primitiva de φ tal que $\inf_x f(x) = 0$. Definimos o tempo-de-paragem

$$\tau = \tau(\mathcal{O}, g) := \inf\{t :\ \text{ocorre (i) } x_t \in \mathcal{O},\ \text{ou (ii) } |x_t| < R \text{ e } \gamma_t \geq g\}.$$

O valor $g \in (0, \bar{\mathrm{g}})$ será especificado em baixo.

Consideramos a sequência $\mathrm{E}_t = \mathrm{E}[f(x_t)\ \mathbb{I}(t < \tau)]$. Introduzimos a notação compacta $f(x_t) =: f_t,\ \ \mathbb{I}(t < \tau) =: I_t,\ \ f'(x_t) =: f'_t = \varphi_t$, e usando que $I_t \le I_{t-1}$, temos

$$\mathrm{E}_t - \mathrm{E}_{t-1} = \mathrm{E}[f_t\, \mathbb{I}_t - f_{t-1}\, \mathbb{I}_{t-1}] \ \le\ \mathrm{E}[(f_t - f_{t-1})\, \mathbb{I}_{t-1}]. \tag{3.21}$$

De seguida, usando a decomposição de Taylor

$$f_t = f(x_{t-1} - \gamma_{t-1} y_t) = f_{t-1} - f'_{t-1}\, \gamma_{t-1} y_t + \frac{1}{2}\, f''(x')\, \gamma_{t-1}^2 y_t^2,$$

sendo x' um ponto entre x_{t-1} e x_t. Substituindo $y_t = \varphi_{t-1} + \xi_t$ e relembrando que $f'_{t-1} = \varphi_{t-1}$ e $f''(x') = \varphi'(x') \le M$, temos

$$f_t - f_{t-1} \le -\gamma_{t-1}\, \varphi_{t-1}(\varphi_{t-1} + \xi_t) + \frac{M}{2}\, \gamma_{t-1}^2\, (\varphi_{t-1} + \xi_t)^2. \tag{3.22}$$

Usando (3.21) e (3.22) e tomando em conta que cada um dos valores $\gamma_{t-1},\ \ \varphi_{t-1},\ \ \mathbb{I}_{t-1}$ é mutuamente independente com ξ_t (ver A1), temos

$$\begin{aligned}
&\mathrm{E}_t - \mathrm{E}_{t-1} \le \mathrm{E}[(-\gamma_{t-1}\, \varphi_{t-1}^2 - \gamma_{t-1}\, \varphi_{t-1}\xi_t + \tfrac{M}{2}\gamma_{t-1}^2\, \varphi_{t-1}^2 + M\gamma_{t-1}^2\, \varphi_{t-1}\xi_t + \tfrac{M}{2}\gamma_{t-1}^2\, \xi_t^2)\, \mathbb{I}_{t-1}] = \\
&= \mathrm{E}[(-\varphi_{t-1}^2 + \tfrac{M}{2}\gamma_{t-1}\, \varphi_{t-1}^2 + \tfrac{M}{2}\gamma_{t-1} S)\gamma_{t-1}\, \mathbb{I}_{t-1}] = \\
&= \mathrm{E}[(-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1} S/2)\gamma_{t-1}\, \mathbb{I}_{t-1}].
\end{aligned} \tag{3.23}$$

Se $\mathbb{I}_{t-1} = 1$ então ou ocorre (i) $x_{t-1} \in [-R, R] \setminus \mathcal{O}$ e $\gamma_{t-1} < g$, ou (ii) $|x_{t-1}| < R$.

No caso (i) temos

$$-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1} S/2 \le -c_0(1 - Mg/2) + MgS/2 =: -c'_g, \tag{3.24}$$

onde $c_0 := \inf\{|\varphi(x)| : x \in [-R, R] \setminus \mathcal{O}\}$; obviamente, $c_0 > 0$. Fixamos $g \in (0, \bar{\mathrm{g}})$ tal que $c'_g > 0$.

No caso (ii), designando $\bar{b}_0 := \inf_{|x| \ge R} \varphi^2(x)$, temos

$$-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1} S/2 \le -\bar{b}_0(1 - M\bar{\mathrm{g}}/2) + M\bar{\mathrm{g}} S/2 =: -c''. \tag{3.25}$$

Usando A6, obtemos que $c'' > 0$.

Denotamos $c = \min\{c'_g, c''\}$. As relações (3.24) e (3.25) implicam que se $\mathbb{I}_{t-1} = 1$ então $-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1} S/2 \le -c < 0$, e por (3.23),

$$\mathrm{E}_t - \mathrm{E}_{t-1} \le -c \cdot \mathrm{E}[\gamma_{t-1}\, \mathbb{I}_{t-1}]. \tag{3.26}$$

Somando ambos os lados de (3.26) sobre $t = 1, \ldots, s$ e denotando $\mathbb{I}_\infty = \mathbb{I}(\tau = \infty) = \min_t \mathbb{I}_t$ obtemos

$$\mathrm{E}_s - \mathrm{E}_0 \leq -c \cdot \mathrm{E}\left[\sum_{i=0}^{s-1} \gamma_i \cdot \mathbb{I}_\infty\right]$$

Temos $\mathrm{E}_s \geq 0$, e x_0 é limitado, de onde $\mathrm{E}_0 < \infty$. Assim, para arbitrários valores de s

$$\mathrm{E}[\sum_{i=0}^{s-1} \gamma_i \cdot \mathbb{I}_\infty] \leq \frac{\mathrm{E}_0}{c} < \infty.$$

Isto implica que *quase-certamente* ou ocorre $\sum_0^\infty \gamma_i < \infty$, ou $\tau = \infty$. O Lema 27 está demonstrado.

□

Fixamos um intervalo aberto $\mathcal{O}$ contendo Z. Denotamos $c_1 := 1 - M\bar{\mathrm{g}}/2$. Relembramos que f é a primitiva de φ tal que $\inf_x f(x) = 0$; a Condição A6 implica que $\lim_{x\to\pm\infty} f(x) = +\infty$. Denotamos $H := \sup_{|x|\leq R} f(x)$. Denotamos também que $c_3 := \bar{\mathrm{g}} \cdot \sup\{|\varphi(x)| : f(x) \leq H\} + 1$, $z^l := \inf\{x : f(x) \leq H\} - c_3$, $z^r := \sup\{x : f(x) \leq H\} + c_3$, $c_2 := \inf\{|\varphi(x)| : x \in [z^l, z^r] \setminus \mathcal{O}\}$, e $\mathrm{K} := \sup\{|\varphi(x)| : x \in [z^l, z^r]\}$. Obviamente que, $c_1 > 0$ e $\mathrm{K} \geq c_2 > 0$.

Seja $g > 0$, $0 < w < 1$. Dizemos que uma sequência determinística (finita ou infinita) $\{z_0, z_1, z_2, \ldots\}$ é (g, w)-admissível se $|z_0| \leq R$ e existem sequências determinísticas $\{q_t\}$, $\{h_t\}$ tais que

1) $|h_t| \leq w$;

2) se $\{z_0, z_1, \ldots, z_t\} \subset [z^l, z^r] \setminus \mathcal{O}$ então $gd^2 \leq q_s \leq \bar{\mathrm{g}}$, $s = 0, 1, \ldots, t$;

3) $z_t = z_{t-1} - q_{t-1}\,\varphi(z_{t-1}) - h_t$, $t = 1, 2, \ldots$.

Proposição 1 *Existem constantes t_0 e w tal que cada sequência $(g, w)-$admissível $\{z_t, t = 0, 1, \ldots, t_0\}$ tem uma intersecção não vazia com $\mathcal{O}$.*

Prova. Seja $w := \min\{1,\, gd^2c_2^2c_1/(2\mathrm{K})\}$. Denotamos $\tilde{t} = \inf\{t : z_t \in \mathcal{O}\}$; $\tilde{t}$ tomando valores de $\{0,\, 1, \ldots, t_0,\, +\infty\}$. Usamos a notação compacta $f_t := f(z_t)$, $f'_t = \varphi_t := \varphi(z_t)$. Temos

$$f_t = f(z_{t-1} - q_{t-1}\varphi_{t-1} - h_t) = f(z_{t-1} - q_{t-1}\varphi_{t-1}) - f'(\tilde{z}).h_t, \tag{3.27}$$

onde $\tilde{z}$ é um ponto entre $z_{t-1} - q_{t-1}\varphi_{t-1}$ e $z_{t-1} - q_{t-1}\varphi_{t-1} - h_t$.

De seguida, temos

$$f(z_{t-1} - q_{t-1}\varphi_{t-1}) = f_{t-1} - f'_{t-1}q_{t-1}\varphi_{t-1} + \frac{1}{2}f''(\hat{z})\, q_{t-1}^2\varphi_{t-1}^2, \tag{3.28}$$

onde $\hat{z}$ é um ponto entre z_{t-1} e $z_{t-1} - q_{t-1}\varphi_{t-1}$.

Vamos provar por indução que

$$\text{se } 0 \le s \le \tilde{t} \text{ então } f_s \le H - s \cdot gd^2c_2^2c_1/2. \tag{3.29}$$

Para $s = 0$, (3.29) segue da condição $|z_0| \le R$ e da definição de H. Agora, seja $1 \le t \le \tilde{t}$; supomos que a fórmula (3.29) é válida para $0 \le s \le t-1$ e provamos isso para $s = t$. Para $0 \le s \le t-1$, temos $f(z_s) \le H$, $z_s \notin \mathcal{O}$, portanto $z_s \in [z^l, z^r] \setminus \mathcal{O}$; assim, por 2), $gd^2 \le q_s \le \bar{\mathrm{g}}$ para $0 \le s \le t-1$. Temos $f(z_{t-1}) \le H$, $|q_{t-1}\varphi_{t-1}| \le \bar{\mathrm{g}} \cdot \sup\{|\varphi(x)| : f(x) \le H\}$, e $|h_t| \le w \le 1$, portanto $|q_{t-1}\varphi_{t-1}| \le c_3$, $|q_{t-1}\varphi_{t-1} + h_t| \le c_3$, e por isso, $z_{t-1} - q_{t-1}\varphi_{t-1} \in [z^l, z^r]$, $z_{t-1} - q_{t-1}\varphi_{t-1} - h_t \in [z^l, z^r]$, logo $\tilde{z}$ também pertence a $[z^l, z^r]$. Isto implica que $|\varphi(\tilde{z})| = |f'(\tilde{z})| \le \mathrm{K}$. Então, combinando (3.27) e (3.28) e usando que $|h_t| \le w$ e $|f''(\hat{z})| = |\varphi'(\hat{z})| \le M$, obtemos

$$f_t \le f_{t-1} - q_{t-1}\varphi_{t-1}^2(1 - \frac{1}{2}q_{t-1}M) + w\mathrm{K}. \tag{3.30}$$

Temos que $z_{t-1} \in [z^l, z^r] \setminus \mathcal{O}$, portanto $|\varphi(z_{t-1})| = |\varphi_{t-1}| \ge c_2$. Usando que $q_{t-1} \ge gd^2$, $1 - \frac{1}{2}q_{t-1}M \ge c_1$, e $w\mathrm{K} \le gd^2c_2^2c_1/2$, obtemos de (3.30) que

$$f_t \le f_{t-1} - gd^2c_2^2c_1/2,$$

e usando a hipótese de indução, concluímos que

$$f_t \le H - t \cdot gd^2c_2^2c_1/2.$$

A fórmula (3.29) está demonstrada.

Seja $t_0 := \lfloor 2H/(gd^2c_2^2c_1) \rfloor + 1$; onde $\lfloor z \rfloor$ significa a parte inteira de z. Então, tomando em conta que $f_s \ge 0$, de (3.29) concluímos que $\tilde{t} < t_0$, assim a Proposição 1 está demonstrada. □

Proposição 2 *Se $\gamma_{t-1} < 1/(3M)$, $|\xi_t| < c_2$, $|\xi_{t+1}| < c_2$, x_{t-1} e x_t pertence a $[z^l, z^r] \setminus \mathcal{O}$, então $\gamma_{t+1} \ge \gamma_t$.*

Prova. Usando a notação $\varphi_t := \varphi(x_t)$, obtemos

$$\varphi_t = \varphi(x_{t-1} - \gamma_{t-1} \cdot (\varphi_{t-1} + \xi_t)) = \varphi_{t-1} - \varphi'(\tilde{x}) \cdot \gamma_{t-1} \cdot (\varphi_{t-1} + \xi_t),$$

onde $\tilde{x}$ é um ponto entre x_{t-1} e x_t. Assim,

$$\varphi_t\varphi_{t-1} = \varphi_{t-1}^2 \cdot [1 - \varphi'(\tilde{x})\gamma_{t-1} \cdot (1 + \xi_t/\varphi_{t-1})].$$

Usando que $|\varphi'(\tilde{x})| \leq M$, $\gamma_{t-1} < 1/(3M)$, $|\xi_t| < c_2$, $|\varphi_{t-1}| \geq c_2$, obtemos $1 - \varphi'(\tilde{x})\gamma_{t-1} \cdot (1 + \xi_t/\varphi_{t-1}) \geq 1/3$, e daqui que $\varphi_{t-1}\varphi_t > 0$. Mais, usando que $|\xi_t| < c_2$, $|\xi_{t+1}| < c_2$, $|\varphi_{t-1}| \geq c_2$, $|\varphi_t| \geq c_2$, obtemos

$$y_t y_{t+1} = \varphi_{t-1}\varphi_t \cdot (1 + \xi_t/\varphi_{t-1}) \cdot (1 + \xi_{t+1}/\varphi_t) > 0.$$

Isto implica que $\gamma_{t+1} = \min\{u\gamma_t, \bar{\mathrm{g}}\} \geq \gamma_t$.

□

Lema 28 *Para cada conjunto aberto $\mathcal{O}$ contendo* Z*, e cada $g > 0$, existe $\delta = \delta(\mathcal{O}, g) > 0$ tal que*

$$\textit{se } |x_0| \leq R, \gamma_0 \geq g \textit{ então } \mathrm{P}(\textit{para alguns } t, x_t \in \mathcal{O}) \geq \delta$$

Prova. Sem perda de generalidade supomos que $g < 1/(3M)$. Definimos o evento

$$A := \{|\xi_i| < \min\{c_2, w/\bar{g}\}\}, \quad i = 1, 2, \ldots, t_0\},$$

onde w e t_0 estão definidos na demonstração da Proposição 1: $w = \min\{1, gd^2c_2^2c_1/(2K)\}$, $t_0 = \lfloor H/(gd^2c_2^2c_1)\rfloor + 1$.

Denotamos

$$\delta := P(A) = (\mathrm{P}(|\xi_1| < \min\{c_2, w/\bar{\mathrm{g}}\}))^{t_0};$$

e por consequência de A3 (a), $\delta > 0$. Vamos ver que para cada evento elementar $\omega \in A$, a sequência $\{z_t = x_t(\omega), t = 0, 1, \ldots, t_0\}$ é $(g, w)-$admissível.

Temos $|z_0| = |x_0(\omega)| < R$. Mais, temos $z_t = z_{t-1} - q_{t-1}\varphi(z_{t-1}) - h_t$, com $q_{t-1} = \gamma_{t-1}(\omega)$, $h_t = \gamma_{t-1}(\omega)\xi_t(\omega)$, e usando que $\gamma_{t-1}(\omega) \leq \bar{\mathrm{g}}$ e $|\xi_t(\omega)| < w/\bar{\mathrm{g}}$, obtemos $|h_t| \leq w$. Assim, as condições 1) e 3) são verificadas.

Agora, seja $\{z_0, z_1, \ldots, z_t\} \subset [z^l, z^r] \setminus \mathcal{O}$, $t \leq t_0$. Seja $s_0 \in \{0, 1, 2, \ldots, t\}$ o menor valor tal que $q_{s_0} = \min\{q_0, q_1, \ldots, q_t\}$. Se $s_0 = 0$ então $\min\{q_0, q_1, \ldots, q_t\} = q_0 = \gamma_0(\omega) \geq g \geq gd^2$. Se $s_0 = 1$ então $\min\{q_0, q_1, \ldots, q_t\} = q_1 = \gamma_1(\omega) \geq \gamma d \geq gd^2$. Se $s_0 \geq 2$ então $\gamma_{s_0-2}(\omega) \geq 1/(3M)$; por outro lado, usando que $|\xi_{s_0-1}| < c_2$, $|\xi_{s_0}| < c_2$, $x_{s_0-2}(\omega)$ e $x_{s_0-1}(\omega)$ pertence a $[z^l, z^r] \setminus \mathcal{O}$, e aplicando a Proposição 2, concluímos que $\gamma_{s_0}(\omega) \geq \gamma_{s_0-1}(\omega)$, o que contradiz a definição de s_0.

Assim, $\gamma_{s_0}(\omega) \geq 1/(3M) \cdot d^2 \geq gd^2$, e por conseguinte, $\min\{q_0, q_1, \ldots, q_t\} = \gamma_{s_0}(\omega) \geq gd^2$. Assim, a condição 2) é também verificada.

Agora, aplicando a Proposição 1 à sequência $(g, w)-$admissível $\{z_t\}$, concluímos que existe um valor não negativo $\tau \leq t_0$ tal que $z_\tau = x_\tau(\omega) \in \mathcal{O}$. Isto implica que

$$\mathrm{P}(\text{para alguns } t, x_t \in \mathcal{O}) \geq \mathrm{P}(A) = \delta.$$

O Lema 28 está demonstrado.

□

Lema 29 *Se $\sum_t \gamma_t = \infty$ então para cada conjunto aberto $\mathcal{O}$ contendo* Z *existe t tal que $x_t \in \mathcal{O}$.*

Prova. Vamos fixar um conjunto aberto $\mathcal{O} \supset Z$ e denotamos $\delta = \delta(\mathcal{O}, g(\mathcal{O}))$. Combinando os Lemas 27 e 28, concluímos que para cada $\mathcal{O} \supset$ Z existe $\delta > 0$ tal que seja quais forem as condições iniciais x_0, γ_0, γ_1,

$$\mathrm{P}(\text{existe } t \text{ tal que } x_t \in \mathcal{O} \Big| \sum_t \gamma_t = \infty) > \delta.$$

Podemos escolher uma função inteira e mensurável $n(\cdot, \cdot, \cdot)$ definida em $\mathbb{R} \times (0, \bar{\mathrm{g}}] \times (0, \bar{\mathrm{g}}]$ tal que para $\nu = n(x_0, \gamma_0, \gamma_1)$ temos

$$\mathrm{P}(\text{para algum } t \leq \nu, x_t \in \mathcal{O} \Big| \sum_t \gamma_t = \infty) > \delta/2$$

Designamos

$$\bar{p} = \sup \mathrm{P}(\text{para cada } t, x_t \notin \mathcal{O} \Big| \sum_t \gamma_t = \infty),$$

o supremo tomado sobre todas as condições iniciais x_0, γ_0, γ_1. Fixando x_0, γ_0, γ_1, então

$$\begin{aligned} &\mathrm{P}(\text{para cada } t, x_t \notin \mathcal{O} \Big| \textstyle\sum_t \gamma_t = \infty) = \\ &= \mathrm{P}(\text{para cada } t > \nu, x_t \notin \mathcal{O} \Big| \text{ para cada } t \leq \nu, x_t \notin \mathcal{O} \text{ e } \textstyle\sum_t \gamma_t = \infty) \cdot \\ &\cdot \mathrm{P}(\text{para cada } t \leq \nu, x_t \notin \mathcal{O} | \textstyle\sum_t \gamma_t = \infty) \leq \bar{p}(1 - \delta/2). \end{aligned} \tag{3.31}$$

Tomando o supremo do lado esquerdo de (3.31) sobre todo o $(x_0, \gamma_0, \gamma_1) \in \mathbb{R} \times (0, \bar{\mathrm{g}}] \times (0, \bar{\mathrm{g}}]$, obtemos $\bar{p} \leq \bar{p}(1 - \delta/2)$, logo $\bar{p} = 0$.

O Lema 29 está demonstrado.

□

Relembramos a definição de L na Condição A3 e denotamos $\mathcal{O}_* = \{x : |\varphi(x)| < L/2\}$.

Lema 30 *Para quaisquer conjuntos abertos $\mathcal{O}$, $\mathcal{O}_1$ tais que $\bar{\mathcal{O}} \subset \mathcal{O}_1 \subset \mathcal{O}_*$ e para cada $w > 0$ existe $\delta = \delta(\mathcal{O}, \mathcal{O}_1, w) > 0$ tal que*

$$\textit{se } \; x_0 \in \mathcal{O} \textit{ então } \; \mathrm{P}(\textit{para alguns } n, \; x_n \in \mathcal{O}_1 \; \textit{e } \gamma_n < w) \geq \delta.$$

(Aqui $\bar{\mathcal{O}}$ é o fecho de $\mathcal{O}$).

Prova. Denotamos $n = \lfloor \frac{\ln \bar{g} - \ln w}{\ln(1/d)} \rfloor + 1$. Denotamos também

$$\varepsilon = \min \left\{ \frac{L}{2},\ \frac{\partial(\mathcal{O},\, \mathbb{R} \setminus \mathcal{O}_1)}{n\bar{g}} \right\},$$

onde $\partial(A, B) := \sup_{x \in A} \inf_{y \in B} |x - y|$ para conjuntos arbitrários $A,\ B$ de números reais. Usando a condição A3 (a), obtemos que existe $\delta_1 > 0$ tal que para cada $x \in \mathcal{O}_1$ e para cada inteiro t,

$$\mathrm{P}\left((-1)^{t-1}\varphi(x) < (-1)^t \xi_1 < (-1)^{t-1}\varphi(x) + \varepsilon\right) \geq \delta_1.$$

Isto implica que se $x_0 \in \mathcal{O}$ então

$$\mathrm{P}(0 < (-1)^t y_t < \varepsilon,\ \partial(x_{t-1},\, \mathcal{O}) < (t-1)\bar{g}\varepsilon,\ t = 1,\, 2, \ldots, n+1) \geq \delta_1^{n+1}.$$

Denotando $\delta = \delta_1^{n+1}$ concluímos que as seguintes proposições (i) e (ii) são válidas com probabilidade pelo menos δ:

(i) $\partial(x_n,\, \mathcal{O}) < n\bar{g}\varepsilon \leq \partial(\mathcal{O},\, \mathbb{R} \setminus \mathcal{O}_1)$, então $x_n \in \mathcal{O}_1$;

(ii) quando $t = 2,\, 3, \ldots, n+1$, temos $y_{t-1} y_t < 0$, e $\gamma_t = d\gamma_{t-1}$, por conseguinte $\gamma_n = d^{n-1}\gamma_1 \leq d^{n-1}\bar{g} < w$.

O Lema 30 está demonstrado.

□

Lema 31 *Se* $\sum_t \gamma_t = \infty$, $\mathcal{O}$ *é um conjunto aberto contendo* Z, *e* $w > 0$ *então para algum* $t,\ x_{t-1} \in \mathcal{O}$ *e* $\gamma_t < w$.

Prova. Sem perda de generalidade, supomos que $\mathcal{O}$ é limitado e $\mathcal{O} \subset \mathcal{O}_*$. Escolhemos um conjunto aberto $\mathcal{O}_1$ tal que $\mathrm{Z} \subset \mathcal{O}_1,\ \bar{\mathcal{O}}_1 \subset \mathcal{O}$; aplicando os Lemas 29 e 30, obtemos que para $\delta = \delta(\mathcal{O}_1, \mathcal{O}, w)$ e para condições arbitrárias,

$$\mathrm{P}(\text{para algum } t,\ x_t \in \mathcal{O} \text{ e } \gamma_t < w) > \delta.$$

Repetindo o argumento do Lema 29, concluímos que existe t tal que $x_t \in \mathcal{O}$ e $\gamma_t < w$.

□

Daqui em diante supomos que $\mathrm{k} > k_+(0)$. Escolhemos k' tal que $k_+(0) < k' < \mathrm{k}$; usando A3 (b), temos que para algum $\varepsilon_0 > 0,\ \mathrm{P}(\xi_1\xi_2 > 0,\ \text{ou } |\xi_1| < \varepsilon_0,\ \text{ou } |\xi_2| < \varepsilon_0) \leq k'$. Denotemos $\mathcal{O}_0 = \{x : |\varphi(x)| < \varepsilon_0\}$ e $\tau = \inf\{t :\ x_t \notin \mathcal{O}_0\}$. Sem perda de generalidade, supomos que $\mathcal{O}_0$ é limitado.

Lema 32 *Supomos que* $\mathrm{k} > k_+(0)$, *então existe uma constante* $b > 0$ *e uma função monótona decrescente* $p(\cdot)$ *tal que* $\lim_{a\to+\infty} p(a) = 0$ *e*

$$\textit{se}\ \ \gamma_0 < w\ \textit{então}\ \ \mathrm{P}(\ln\gamma_t < \ln v - bt\ \textit{para cada}\ t < \tau) > 1 - p(v/w).$$

Prova. Definimos as sequências $\{\rho_t\}$ e $\{\sigma_t\}$ por

$$\begin{aligned}\rho_t &= \ln u \cdot \mathbb{I}(\xi_{t-1}\xi_t > 0,\ \text{ou}\ |\xi_{t-1}| < \varepsilon_0,\ \text{ou}\ |\xi_t| < \varepsilon_0) + \\ &+ \ln d \cdot \mathbb{I}(\xi_{t-1}\xi_t \le 0\ \text{e}\ |\xi_{t-1}| \ge \varepsilon_0\ \text{e}\ |\xi_t| \ge \varepsilon_0),\end{aligned}$$

$$\sigma_t = \ln w + \sum_{i=1}^{t} \rho_i.$$

Usando (3.5) e a definição de τ, obtemos que para cada $t < \tau,\ \gamma_t \le \sigma_t$. As variáveis ρ_t são identicamente distribuídas. Tomando os valores $\ln u$ e $\ln d$,

$$\begin{aligned}\mathrm{E}\rho_t &= \ln u \cdot \mathrm{P}(\xi_{t-1}\xi_t > 0,\ \text{ou}\ |\xi_{t-1}| < \varepsilon_0,\ \text{ou}\ |\xi_t| < \varepsilon_0) + \\ &+ \ln d \cdot \mathrm{P}(\xi_{t-1}\xi_t \le 0\ \text{e}\ |\xi_{t-1}| \ge \varepsilon_0\ \ \text{e}\ |\xi_t| \ge \varepsilon_0) \le \\ &\le \ln u \cdot k' + \ln d \cdot (1 - k') < \ln u \cdot \mathrm{k} + \ln d \cdot (1 - \mathrm{k}) = 0.\end{aligned}$$

Mais, as variáveis no conjunto $\{\rho_t,\ t\ \text{par}\}$, e também as variáveis no conjunto $\{\rho_t,\ t\ \text{ímpar}\}$, são independentes.

Denotamos $b = -\mathrm{E}\rho_t/2$. Temos

$$\mathrm{P}(\ln\gamma_t < \ln v - bt\ \ \text{para cada}\ t < \tau) \ge \mathrm{P}(\sigma_t < \ln v - bt\ \ \text{para cada}\ t) =$$

$$= \mathrm{P}(\sum_{i=1}^{t}(\rho_i + 2b) < \ln v - \ln w + bt\ \ \text{para cada}\ t) \ge 1 - p(v/w),$$

onde $p(a) = p_1(a) + p_2(a)$,

$$p_1(a) = \mathrm{P}\left(\sum_{1\le i\le t}{}'(\rho_i + 2b) \ge \frac{\ln a}{2} + \frac{b}{2}t\ \ \text{para cada}\ t\right),$$

$$p_2(a) = \mathrm{P}\left(\sum_{1\le i\le t}{}''(\rho_i + 2b) \ge \frac{\ln a}{2} + \frac{b}{2}t\ \ \text{para cada}\ t\right);$$

a soma $\sum'$ ($\sum''$) é tomada sobre todos os valores pares (ímpares) de i. Ambas as somas $\sum'$ e $\sum''$ são somas de v.a. i.i.d. com média zero, e assim ambos $p_1(a)$ e $p_2(a)$ tendem para zero quando $a \to +\infty$. O Lema 32 está demonstrado.

□

Definimos os tempos de paragem $\tau_v = \inf\{t :\ x_t \notin \mathcal{O}_0\ \text{ou}\ \ln\gamma_t \ge \ln v - bt\}$. Relembramos que f é a primitiva de φ tal que $\inf_x f(x) = 0$. Fixamos um intervalo aberto $\mathcal{O}'$ tal que $\mathrm{Z} \subset \mathcal{O}' \subset \mathcal{O}_0$ e $\sup_{x\in\mathcal{O}'} f(x) < \inf_{x\notin\mathcal{O}_0} f(x)$, e denotamos $\delta = \inf_{x\notin\mathcal{O}_0} f(x) - \sup_{x\in\mathcal{O}'} f(x)$.

Lema 33 *Seja* $\mathrm{k} > k_+(0)$, $x_0 \in \mathcal{O}'$, *e* $\gamma_0 < w$, *então*

$$\mathrm{P}(\tau_v < \infty) \leq K\, v^2 + p(v/w);$$

aqui K *é uma constante positiva, e* $p(\cdot)$ *satisfaz o enunciado do Lema 32.*

Prova. Usaremos a notação do Lema 27: $f_t := f(x_t)$ e $\varphi_t := \varphi(x_t)$. De acordo com (3.22), temos

$$f_t - f_{t-1} \leq -\gamma_{t-1}\varphi_{t-1}(\varphi_{t-1} + \xi_t) + \frac{M}{2}\,\gamma_{t-1}^2(\varphi_{t-1} + \xi_t)^2 \leq$$
$$\leq -\gamma_{t-1}\varphi_{t-1}\xi_t + M\gamma_{t-1}^2(\varphi_{t-1}^2 + \xi_t^2).$$

Isto implica que $f_t - f_1 \leq Q'_t + Q''_t$, com

$$Q'_t = \big|\sum_{i=2}^{t} \gamma_{i-1}\varphi_{i-1}\xi_i\big|, \qquad Q''_t = M\sum_{i=2}^{t} \gamma_{i-1}^2(\varphi_{i-1}^2 + \xi_i^2).$$

Usando o Lema 32, temos

$$\mathrm{P}(\tau_v < \infty) \;\leq\; p(v/w) + P' + P'',$$

onde

$$P' = \mathrm{P}(Q'_{\tau_v} \geq \delta/2) \quad \text{e} \quad P'' = \mathrm{P}(Q''_{\tau_v} \geq \delta/2).$$

De acordo com a desigualdade de Chebyshev,

$$P' \;\leq\; \frac{4}{\delta^2}\, EQ'^2_{\tau_v} = \frac{4}{\delta^2}\sum_{i,j=1}^{\infty} E_{ij},$$

onde

$$E_{ij} \;=\; E\left[\gamma_{i-1}\varphi_{i-1}\xi_i \;\mathbb{I}(i-1 < \tau_v)\cdot \gamma_{j-1}\varphi_{j-1}\xi_j \;\mathbb{I}(j-1 < \tau_v)\right].$$

usando que os valores $\gamma_i,\ \varphi_i,\ \xi_i$, e $\mathbb{I}(i < \tau_v)$ são $\mathcal{F}_i$-mensuráveis, e com as Condições A1 e A2, obtemos que para $i \neq j,\ E_{ij} = 0$, e para $i = j$,

$$E_{ii} = E\left[\gamma_{i-1}^2\varphi_{i-1}^2\,\mathbb{I}(i-1 < \tau_v)\cdot\xi_i^2\right] \leq v^2 e^{-2bi}\sup_{x\in\mathcal{O}_0}\varphi^2(x)\cdot S.$$

Assim,

$$P' \;\leq\; \frac{4}{\delta^2}\sum_{i=2}^{\infty} E_{ii} \leq \frac{4v^2 S}{\delta^2}\;\frac{e^{-4b}}{1-e^{-2b}}\;\sup_{x\in\mathcal{O}_0}\varphi^2(x).$$

De forma semelhante,

$$P'' \;\leq\; \frac{2}{\delta}\, EQ''_{\tau_v} = \frac{2M}{\delta}\sum_{i=2}^{\infty} E\left[\gamma_{i-1}^2(\varphi_{i-1}^2 + \xi_i^2)\,\mathbb{I}(i-1 < \tau_v)\right] \leq$$

$$\leq \frac{2Mv^2}{\delta}\sum_{i=2}^{\infty} e^{-2bi}\left(\sup_{x\in\mathcal{O}_0}\varphi^2(x)+S\right) = \frac{2Mv^2}{\delta}\,\frac{e^{-4b}}{1-e^{-2b}}\left(\sup_{x\in\mathcal{O}_0}\varphi^2(x)+S\right).$$

Tomando

$$K \;=\; \left[\frac{4S}{\delta^2}\sup_{x\in\mathcal{O}_0}\varphi^2(x) \;+\; \frac{2M}{\delta}\left(\sup_{x\in\mathcal{O}_0}\varphi^2(x)+S\right)\right]\frac{e^{-4b}}{1-e^{-2b}},$$

obtemos que $P' + P'' \leq K\,v^2$. O Lema 33 está demonstrado.

□

Lema 34 *Se* $\mathrm{k} > k_+(0)$ *então* $\sum_t \gamma_t < \infty$.

Prova. Da definição de τ_v facilmente verificamos que se $\tau_v = \infty$ para alguns $v > 0$, então $\sum_t \gamma_t < \infty$. Tal implica que para cada $v > 0$

$$\mathrm{P}\left(\sum \gamma_t \;=\; \infty\right) \leq \mathrm{P}(\tau_v < \infty). \tag{3.32}$$

Em virtude do Lema 33, se $x_0 \in \mathcal{O}'$ e $\gamma_0 < w$ então

$$\mathrm{P}(\tau_{\sqrt{w}} \;<\; \infty) \;\leq\; Kw + p(1/\sqrt{w}). \tag{3.33}$$

Combinando (3.32) e (3.33), obtemos que para cada $w > 0$

$$\mathrm{P}\left(\sum \gamma_t = \infty \mid x_0 \in \mathcal{O}' \text{ e } \gamma_0 < w\right) \;\leq\; Kw \;+\; p(1/\sqrt{w}). \tag{3.34}$$

Definimos o evento $\mathcal{A}_w = \{\text{para cada } t,\ x_t \in \mathcal{O}' \text{ e } \gamma_t < w\}$, assim, por (3.34),

$$\mathrm{P}\left(\sum \gamma_t = \infty \;\middle|\; \mathcal{A}_w\right) \leq Kw + p(1/\sqrt{w}). \tag{3.35}$$

Denotamos por $\bar{\mathcal{A}}_w$ o evento complementar, $\bar{\mathcal{A}}_w = \{\text{para cada } t,\ x_t \notin \mathcal{O}' \text{ ou } \gamma_t \geq w\}$. Em virtude do Lema 31,

$$\mathrm{P}\left(\sum \gamma_t = \infty \text{ e } \bar{\mathcal{A}}_w\right) = 0. \tag{3.36}$$

Usando (3.35) e (3.36), obtemos

$$\mathrm{P}\left(\sum \gamma_t = \infty\right) = \mathrm{P}\left(\sum \gamma_t = \infty \text{ e } \mathcal{A}_w\right) + \mathrm{P}\left(\sum \gamma_t = \infty \text{ e } \bar{\mathcal{A}}_w\right) \leq$$

$$\leq (Kw + p(1/\sqrt{w})) \cdot \mathrm{P}(\mathcal{A}_w).$$

Tomando em conta que w pode ser escolhido arbitrariamente pequeno e que $Kw + p(1/\sqrt{w}) \to 0$ quando $w \to 0^+$, concluímos que $\mathrm{P}\left(\sum_t \gamma_t = \infty\right) = 0$.

□

Finalmente, estamos em condições de demonstrar o Teorema. Supomos que $\mathrm{k} < \inf_z k_-(z)$, então $V_-^{[\mathrm{k}]} = \emptyset$, e pelo Lema 26, $\{x_t\}$ diverge. Assim, o enunciado (b) do Teorema está demonstrado.

Por outro lado, de acordo com o Lema 34, se k $> k_+(0)$ então $\sum_t \gamma_t < \infty$, e pelos Lemas 25 e 26, a sequência $\{x_t\}$ converge para um ponto de $V_-^{[\mathrm{k}]}$. Assim, o enunciado (a) do Teorema está também demonstrado. □

Capítulo 4

Estudos Numéricos

Apresentamos estudos numéricos dos métodos referidos na tese para o caso duma função unidimensional, o caso de um campo de vectores originado pela função de Rosenbrock e ainda uma breve referência a uma rede neuronal para aprendizagem da função lógica ou-exclusivo.

O estudo foi realizado comparando o comportamento dos algoritmos: Robbins-Monroe (RM) e Kesten (K), e ainda os algoritmos propostos, Kesten generalizado (Kg) e passo multiplicativo (Mul).

4.1 Caso unidimensional

A inspiração para a função $\varphi : \mathbb{R} \to \mathbb{R}$ escolhida para as simulações proveio das redes neuronais para as quais ocorrem na superfície do erro quadrático patamares com pouca inclinação em regiões afastadas duma possível solução. A função usada foi

$$\varphi(x) = \text{sen}(\alpha \tanh(x)) \tag{4.1}$$

em que seleccionámos a constante $\alpha = 19 \times \pi/20$. O esboço da função encontra-se na Figura 4.1. A função tem assímptotas horizontais para a direita ($y \simeq +0.15$) e para a esquerda ($y \simeq -0.15$). O valor da constante α é intencionalmente próximo de π para que as assímptotas sejam próximas ao eixo xx. O declive no zero é $\varphi'(0) = \alpha$.

Nos estudos efectuados houve a preocupação para que os algoritmos Robbins-Monroe, Kesten e Kesten generalizado fossem assimptoticamente semelhantes. Por essa razão consideramos uma forma geral do passo comum a estes algoritmos

$$\gamma_t = \frac{\bar{\gamma}}{E_0 s_t + 1}$$

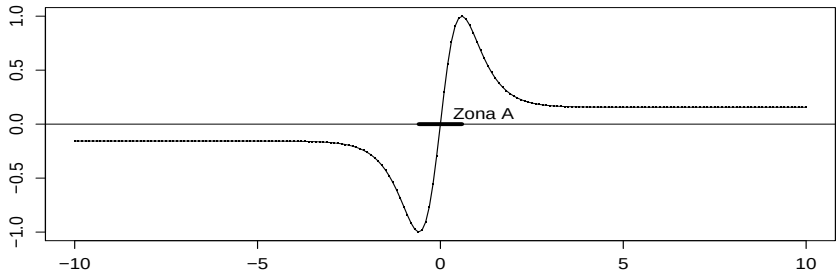


Figura 4.1: Esboço de $\varphi(x) = \text{sen}(19 \times \pi/20 \tanh(x))$.

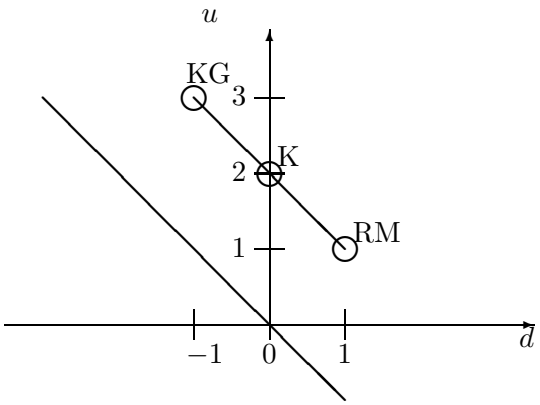


Figura 4.2: Valores para (u, d).

em que $\bar{\gamma}$ é o passo máximo, E_0 é uma constante positiva, $s_0 = 0$, $s_t = [s_{t-1} + \delta(y_{t-1}y_t)]^+$, $t = 1, 2, \ldots$, e onde $[a]^+ = \max\{0, a\}$, e $\delta(x)$ é uma função degrau semelhante à descrita na Figura 2.1 (casos (a), (b) e (c), segundo Capítulo, pág. 17). Definimos que quando $y_{t-1}y_t \leq 0$ então s_t é incrementado por u, uma constante positiva; quando $y_{t-1}y_t > 0$, é adicionado a s_t a uma constante d real, sempre verificando $u > -d$ para que haja convergência (a recta $u+d = 0$ está indicada na Figura 4.2). A Figura 4.2 resume de antemão as escolhas feitas para que haja semelhança do comportamento assimptótico nos três métodos. Para tal, devemos verificar que $\text{E}\{\Delta s_t\} = 1$ quando $\varphi(x_t) \simeq 0$, pois nesta situação o sinal de y_t é determinado apenas por ξ_t ($y_t \simeq \xi_t$). As perturbações $\{\xi_t\}$ provêm duma distribuição normal.

De seguida, apresentamos a lista dos métodos usados e configurações escolhidas:

Algoritmo 1 (Robbins-Monroe, pág. 6) O passo usado teve a forma

$$\gamma_t = \frac{\bar{\gamma}}{E_0 \cdot (t-1) + 1},$$

$t = 1, 2, \ldots$. Esta forma do passo permite o decréscimo do passo a um ritmo dado por E_0 começando no passo máximo $\bar{\gamma}$. O algoritmo de Robbins-Monroe é um caso da generalização introduzida na tese em que $u = d = 1$ (Figura 4.2).

Algoritmo 3 (Kesten, pág. 9) O passo teve a forma

$$\gamma_t = \frac{\bar{\gamma}}{E_0 \cdot s_t + 1}$$

em que $s_0 = 0$, $s_t = s_{t-1} + 2\ \mathbb{I}(y_{t-1}y_t \leq 0)$, $t = 1, 2, \ldots$. O algoritmo de Kesten é um caso da generalização introduzida na tese em que $u = 1$, $d = 0$ mas para que se mantenha a mesma constante assimptótica E_0 de variação do passo devemos escolher $u = 2$, $d = 0$ pois por cada vez que o passo se mantém deverá depois descer 2 unidades para que em média o decréscimo seja equilibrado: supondo que x_t está próximo de zero então $y_t \simeq \xi_t$ pelo que

$$\mathrm{E}\{\Delta s_t\} = 2 \cdot \mathrm{P}(\xi_{t-1}\xi_t \leq 0) = 1.$$

Algoritmo 6 (Kesten generalizado, pág. 6) A forma do passo foi

$$\gamma_t = \frac{\bar{\gamma}}{E_0 \cdot s_t + 1},$$

$t = 1, 2, \ldots$ com $s_0 = 0$ e

$$s_t = [s_{t-1} + 3\ \mathbb{I}(y_{t-1}y_t \leq 0) + (-1)\ \mathbb{I}(y_{t-1}y_t > 0)]^+,$$

onde $[x]^+ = \max\{0, x\}$.

A Figura 4.2 ilustra as escolhas $(u, d) = (1, 1)$ para RM e $(u, d) = (2, 0)$ para K. Desta figura, por analogia, foi escolhido o par $(u, d) = (3, -1)$ que mantém a constante E_0 como taxa assimptótica de variação de s_t. Novamente, supondo x_t próximo de zero temos $y_t \simeq \xi_t$ e

$$\mathrm{E}\{\Delta s_t\} = 3\mathrm{P}(\xi_{t-1}\xi_t \leq 0) + (-1)\mathrm{P}(\xi_{t-1}\xi_t > 0) = 1.$$

Algoritmo 7 (Passo multiplicativo, pág. 11) O passo é dado por

$$\gamma_t = \begin{cases} \min\{u\gamma_{t-1}, \bar{\gamma}\} & \text{se } y_{t-1}y_t > 0, \\ d\gamma_{t-1} & \text{se } y_{t-1}y_t \leq 0, \end{cases} \quad t = 2, 3, \ldots$$

com $0 < d < 1 < u$, em que escolhemos $\gamma_0 = \bar{\gamma}$.

O passo máximo $\bar{\gamma} = 2/\alpha$, escolhido para todos os métodos, satisfaz a condição A.3 de convergência (Capítulo 2) pois $\sup_x |\varphi'(x)| = \alpha$.

Para os algoritmos Robbins-Monroe, Kesten e Kesten generalizado, foram realizados estudos para valores de E_0 no conjunto:

$$\{2\alpha, 3\alpha/2, \alpha, \alpha/2, \alpha/20, \alpha/200\}.$$

Estes valores estão em torno do declive $\varphi'(0) = \alpha$ porque $E_0 = \alpha$ garante, pela teoria, trajectórias mais eficientes em média quando x_t está próximo da solução. Para $E_0 = 2\alpha$ o passo γ_t tenderá a decrescer mais rápido que para $E_0 = \alpha/200$.

Para o algoritmo de passo multiplicativo foi fixado o parâmetro $u = 1.1$ (sem razão especial) e determinados valores para o parâmetro d da igualdade $ud = c$ onde

$$c \in \{0.9, 0.95, 0.99, 0.995, 0.999, 1, 1.01, 1.03, 1.05, 1.07, 1.09\}$$

Relembramos, que pela teoria, para valores $ud < 1$ o algoritmo converge e para valores $ud > 1$ o algoritmo diverge e que a teoria não especifica o que ocorre para $ud = 1$.

O comportamento dos algoritmos e configurações foram analisados na resolução dos seguintes problemas, todos usando a mesma função φ definida em (4.1), sendo as perturbações das medidas obtidas duma distribuição normal com desvio S a definir:

$x_0 = 0, S = 1$**:** Parte-se da solução $x^* = 0$ com perturbações elevadas em cada medida da função $\varphi(x)$. Pretende-se observar se os métodos permanecem na solução.

$x_0 = 5, S = 1$**:** Parte-se longe da solução onde $\varphi(x) \simeq \pm 0.15$. As perturbações elevadas afectam o sinal da medida perturbada de $\varphi(x)$ sendo muito frequente que o sinal medido seja o de ξ_t e não o de $\varphi(x_t)$. Este problema é difícil pois os sinais de y_t são bastante semelhantes ao que se observa no zero da função φ.

$x_0 = 10, S = 0.1$**:** Parte-se longe da solução mas com um perturbação pequena onde a maioria das medidas perturbadas $y = \varphi(x) + \xi$ manterá o sinal de $\varphi(x)$, correcto. Foi esta situação que motivou a generalização do algoritmo de Kesten.

No estudo numérico foram realizadas 40 experiências para cada método e realizadas 20 000 iterações. Os resultados foram compilados em gráficos de trajectórias médias e em histogramas do valor de x_t obtido na última iteração. As Figuras 4.4 a 4.18 contém estes elementos por cada problema e algoritmo para as várias configurações. Da análise destes gráficos resultou a

Tabela 4.1 que resume os comportamentos médios por cada problema e da qual se tiraram as seguintes conclusões baseadas na seguinte classificação de trajectórias:

óptimo Na última iteração a trajectória é semelhante à óptima, que ocorreu com $x_0 = 0$ e $E_0 = \alpha$, ou seja $\log_{10} x_T^2 < -2$. Um algoritmo e configuração que produza uma trajectória nestas circunstâncias é classificadas de '*óptimo*'.

satisfatório Um algoritmo e configuração que produza trajectórias que sejam um pouco piores que o definido acima mas que verifiquem $-2 < \log_{10} x_T^2 < -1$ na última iteração é designado de '*satisfatório*'.

mau Os algoritmos e configurações que produzam outras trajectórias médias que não as de cima listadas são registados como '*mau*'. Nestes incluem-se os que produzem trajectórias que entram na 'região A' no gráfico muito tardiamente como ocorre para $E_0 = \alpha/40$ por exemplo em alguns casos.

Seguem-se as conclusões.

1. A Tabela 4.1 não permite identificar um único algoritmo adequado e óptimo a todos os problemas estudados. Para usar o que de melhor tem este conjunto de algoritmos convém conhecer o melhor possível o problema em causa. No entanto, nos itens seguintes, vamos resumir algumas considerações gerais sobre cada um dos algoritmos.

2. O algoritmo com passo multiplicativo oferece para valores $ud = 1$ e próximos um meio de alcançar uma boa vizinhança do zero sob condições muito gerais (linhas 10, 11 e 12). Este resultado numérico, já observado em estudos de Almeida et al [1] (usando o gradiente) e em Salomon e Hemmen [43] (usando a variação da função objectivo), revela que o comportamento deste algoritmo é mais rico que aquilo que a teoria prevê neste trabalho. Para outros valores de ud as soluções encontradas podem não ser boas (linhas 10 e 11).

3. O algoritmo de Kesten generalizado teve sucesso para o problema para o qual foi desenvolvido: quando se começa duma posição inicial em que o ruído é pequeno o algoritmo fornece uma rápida aproximação ao zero. Nestas circunstâncias, conhecendo-se o declive no zero de φ, pode ser definido um rápido algoritmo numa fase de transição, e ainda, com comportamento de eficiência máxima assimptótica (linha 9).

Algoritmo		problema	óptimo	satisfatório	mau
Tipo RM	RM	$x_0 = 0$, $S = 1$ (Fig. 4.4, p. 98)	α, $\alpha/2$	2α, $3\alpha/2$, $\alpha/20$, $\alpha/30$, $\alpha/40$, $\alpha/200$	—
		$x_0 = 5$, $S = 1$ (Fig. 4.5, p. 99)	—	$\alpha/200$	2α, $3\alpha/2$, α, $\alpha/2$, $\alpha/20$, $\alpha/30$, $\alpha/40$
		$x_0 = 10$, $S = 0.1$ (Fig. 4.6, p. 100)	—	$\alpha/200$	2α, $3\alpha/2$, α, $\alpha/2$, $\alpha/20$, $\alpha/30$, $\alpha/40$
	Kesten	$x_0 = 0$, $S = 1$ (Fig. 4.7, p. 101)	α, $\alpha/2$	2α, $3\alpha/2$, $\alpha/20$, $\alpha/30$, $\alpha/40$, $\alpha/200$	—
		$x_0 = 5$, $S = 1$ (Fig. 4.8, p. 102)	—	$\alpha/200$	2α, $3\alpha/2$, α, $\alpha/2$, $\alpha/20$, $\alpha/30$, $\alpha/40$
		$x_0 = 10$, $S = 0.1$ (Fig. 4.9, p. 103)	—	$\alpha/20$, $\alpha/30$, $\alpha/40$, $\alpha/200$	2α, $3\alpha/2$, α, $\alpha/2$
	K.g. $u = 3$ $d = -1$	$x_0 = 0$, $S = 1$ (Fig. 4.10, p. 104)	α, $\alpha/2$	$\alpha/20$, $\alpha/30$, $\alpha/40$, $\alpha/200$	2α[**1**], $3\alpha/2$ [**1**]
		$x_0 = 5$, $S = 1$ (Fig. 4.11, p. 105)	—	$\alpha/200$	2α, $3\alpha/2$, α, $\alpha/2$, $\alpha/20$, $\alpha/30$, $\alpha/40$
		$x_0 = 10$, $S = 0.1$ (Fig. 4.12, p. 106)	$3\alpha/2$, α, $\alpha/2$	$\alpha/20$, $\alpha/30$, $\alpha/40$, $\alpha/200$	—
Multi-plica-tivo	Vários ud	$x_0 = 0$, $S = 1$ (Fig.4.13-4.14,p.107-108)	—	0.995, 0.999, 1, 1.01	0.90[**2**], 0.95[**2**], 0.99[**2**], 1.03, 1.05, 1.07, 1.09
		$x_0 = 5$, $S = 1$ (Fig. 4.15-4.16, p. 109-110)	—	0.999, 1, 1.01	0.90, 0.95, 0.99, 0.995, 1.03, 1.05, 1.07, 1.09
		$x_0 = 10$, $S = 0.1$ (Fig.i 4.17-4.18, p. 111-112)	—	0.90, 0.95, 0.99, 0.995, 0.999, 1, 1.01, 1.03, 1.05, 1.07, 1.09	—

Notas: [**1**] Ver ponto 3. nos comentários a esta tabela. [**2**] O histograma apresenta ocorrência de amostras cuja trajectória foi má causando fraca eficiência em média, revelando-se assim uma configuação pouco robusta.

Tabela 4.1: Robustez e eficiência das soluções produzidas por cada algoritmo e sua configuração.

Foi observado que em poucas trajectórias partindo de $x_0 = 0$, e usando $u = 3$, $d = -1$ não houve uma aproximação efectiva para o zero nas 20 000 iterações. Experiências numéricas análogas mas com os valores

$$u = 2.5, \quad d = -0.5$$

resolveram estes casos mantendo as mesmas conclusões nos problemas $x_0 = 5, S = 1$ e $x_0 = 10, S = 0.1$.

4. Os algoritmos de Robbins-Monroe e Kesten revelaram-se semelhantes (linhas 1 a 6): máxima eficiência para $E_0 = \alpha$ (e quase máxima para $\alpha/2$).

5. Para o valor $E_0 = \alpha/200$ foi sempre possível obter uma boa trajectória média em qualquer método RM, K, ou Kg. Tal deve-se a que o decréscimo do passo seja suficientemente lento para que o algoritmo consiga atingir a uma pequena vizinhança do zero de φ. Este resultado mostra que o valor de E_0 deve ser escolhido suficientemente pequeno por forma a permitir aos algoritmos o tempo necessário para alcançar o óptimo. Mas dois problemas se levantam: se o valor E_0 escolhido for demasiado baixo perde-se velocidade de convergência quando a trajectória se encontra numa pequena vizinhança do óptimo, e, só se sabe se este valor é baixo conhecendo $\varphi(x^*)$, o que nem sempre é possível.

   Neste sentido, o algoritmo de passo multiplicativo parece oferecer um meio de alcançar os mesmos resultados satisfatórios para qualquer problema mas sem se conhecer a derivada $\varphi'(x^*)$.

4.2 Caso bidimensional

A função de Rosenbrock $(x, y) \mapsto 100(y - x^2)^2 + (x - 1)^2$ tem o seguinte gradiente

$$\varphi(x, y) = \left(-2 + 2x - 400(y - x^2)x, 200y - 200x^2\right) \tag{4.2}$$

e tem sido usada em numerosos testes numéricos. O esboço da função por curvas de nível está na Figura 4.3.

As experiências realizadas seguiram a secção anterior: a mesma expressão de adaptação do passo nos algoritmos de Robbins-Monroe, Kesten e Kesten generalizado mas agora com constante assimptótica E_0 diferente. No passo multiplicativo manteve-se a escolha das constantes (u, d). As experiências obedeceram a:

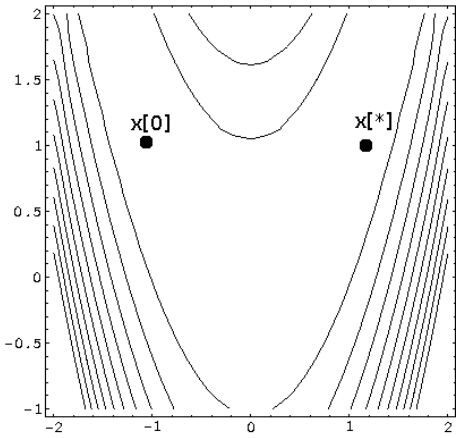


Figura 4.3: Curvas de nível da função de Rosenbrock em que o mínimo ocorre em $(1, 1)$.

- Cada medida de $\varphi(x)$ foi observada com o erro normalmente distribuído $\mathcal{N}(0, 1)$ e independente em cada coordenada.
- O sinal de $y_{t-1} \cdot y_t$, em que $\cdot$ é o produto interno, foi usado para decidir as variações no passo.
- Foram realizadas 40 experiências para cada configuração de 40 000 iterações cada.
- Para a constante E_0 foram usados os valores $\{10, 1, 0.1, 0.01, 0.001, 0.0001\}$. Para valores inferiores a 0.0001 o passo permanece praticamente constante e para valores superiores a 10 as trajectórias não são úteis, como se verá no resumo da experiências.
- O passo multiplicativo usou os mesmos parâmetros que no caso unidimensional: $u = 1.1$ (escolha arbitrária) e d obtido da igualdade $ud = c$ onde

$$c \in \{0.9, 0.95, 0.99, 0.995, 0.999, 1, 1.01., 1.03, 1.05, 1.07, 1.09\}.$$

Os problemas foram:

$x_0 = (1, 1)$, $S = I$: Partindo da solução x^* com perturbações elevadas em cada medida da função $\varphi(x)$. Pretende-se observar se os métodos permanecem numa pequena vizinhança da solução após a primeira iteração.

$x_0 = (-1, 1)$, $S = I$: Partindo dum ponto afastado da solução, pretende-se avaliar a eficácia dos métodos no percurso até ao óptimo. O vale em forma de quadrática é pouco inclinado mas com inclinações acentuadas longe deste. Pretende-se averiguar se os métodos que usam a mudança de sinal de $y_{t-1} \cdot y_t$ conseguem lidar com a curvatura.

As Figuras 4.19 a 4.28 resumem as experiências numéricas realizadas no gráfico das trajectórias médias e no histograma da última iteração e da análise desses gráficos seguem os comentários sobre a aplicação destes algoritmos:

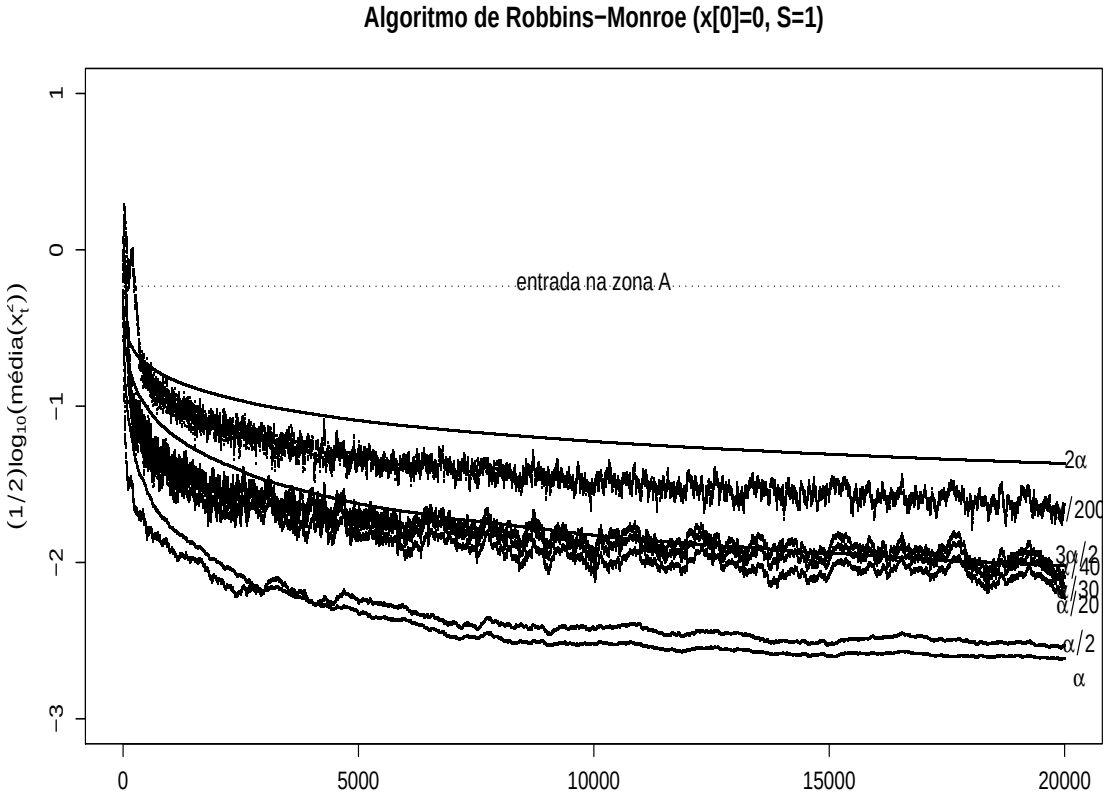
1. Partindo do óptimo a primeira iteração vai afastar-se desse ponto. Nenhum algoritmo conseguiu produzir uma trajectória média que melhorasse, pelo contrário, houve forte pioria das soluções finais encontradas.

2. Partindo do ponto $(-1, 1)$, apenas os algoritmos que permitem passo elevado conseguem produzir uma sequência útil. O algoritmo com constante $E_0 = 10^{-4}$ praticamente não reduz o passo e por isso a trajectória consegue alcançar o zero do campo de vectores. Esta conclusão é válida para os algoritmos Robbins-Monroe, Kesten e Kesten generalizado. Para o passo multiplicativo, os casos $ud > 1$ produzem sequências que mantém o passo suficientemente elevado para que a trajectória atinja uma pequena vizinhança do óptimo embora depois não haja convergência.

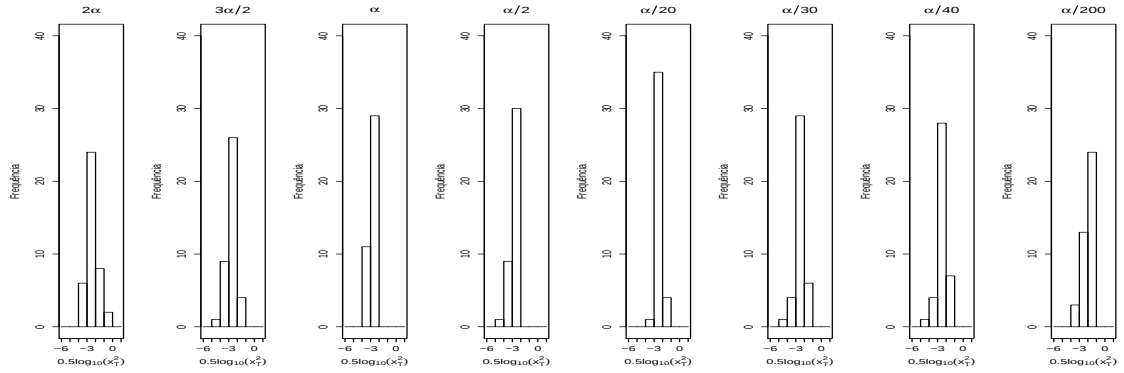
Os resultados desta Secção não encorajam uma aplicação directa destes métodos às redes neuronais mas mesmo assim foi tentado o treino de uma rede neuronal para aprendizagem da função lógica ou-exclusivo. Porém, os mesmos maus resultados foram obtidos: só para situações em que o passo podia permanecer elevado, praticamente igual ao passo máximo, é que houve progressão significativa do treino. Para muitas configurações de parâmetros (u, d), em ambos os algoritmos Kg e Mul, o número máximo de iterações foi atingido sem a rede ter aprendido.

Seguem-se os gráficos das trajectórias médias e histogramas da solução encontrada na última iteração que serviram de base aos comentários do presente Capítulo.

Os resultados numéricos e gráficos foram obtidos programando na linguagem `R` [37]. A rede neuronal acima referida foi programada com `GNU C++` (`gcc.gnu.org`). A Tese foi escrita usando o pacote `MIKTEX` (`www.miktex.de`) tendo sido o livro *The LaTeX Companion* [18] o principal guia de LaTeX.

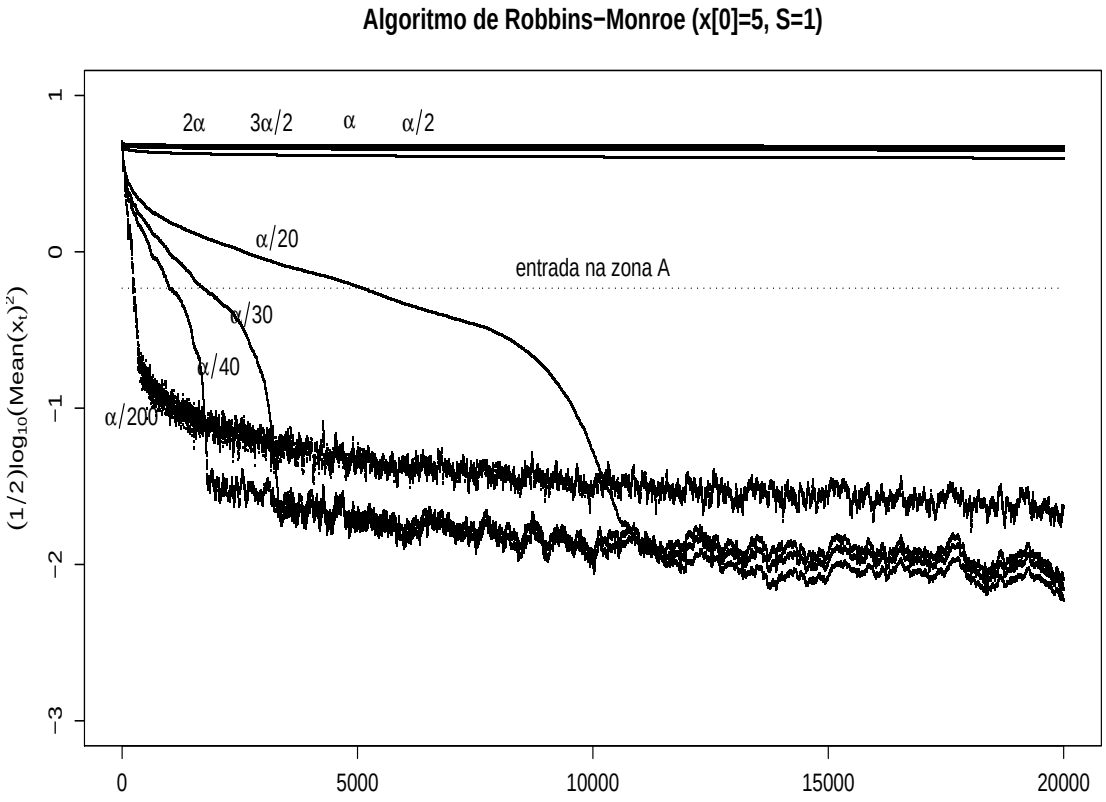


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

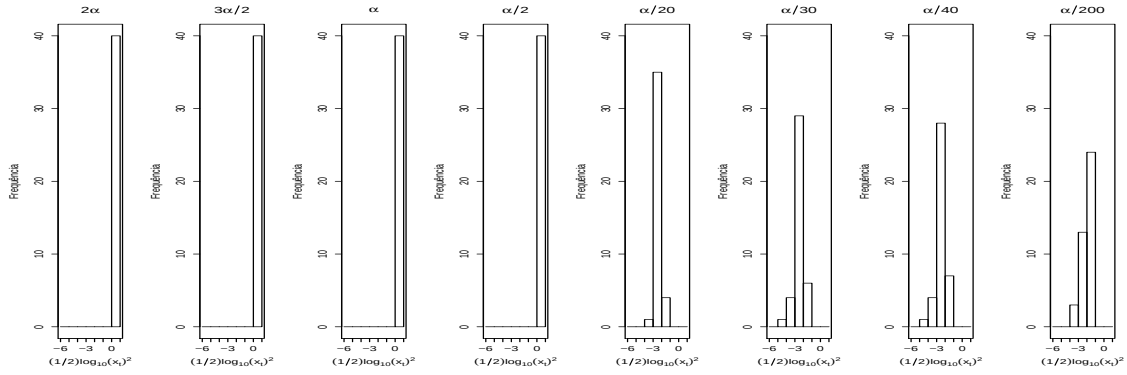


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.4: Descrição estatística da aplicação de várias configurações do algoritmo de Robbins-Monroe na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 0$ (zero da função).

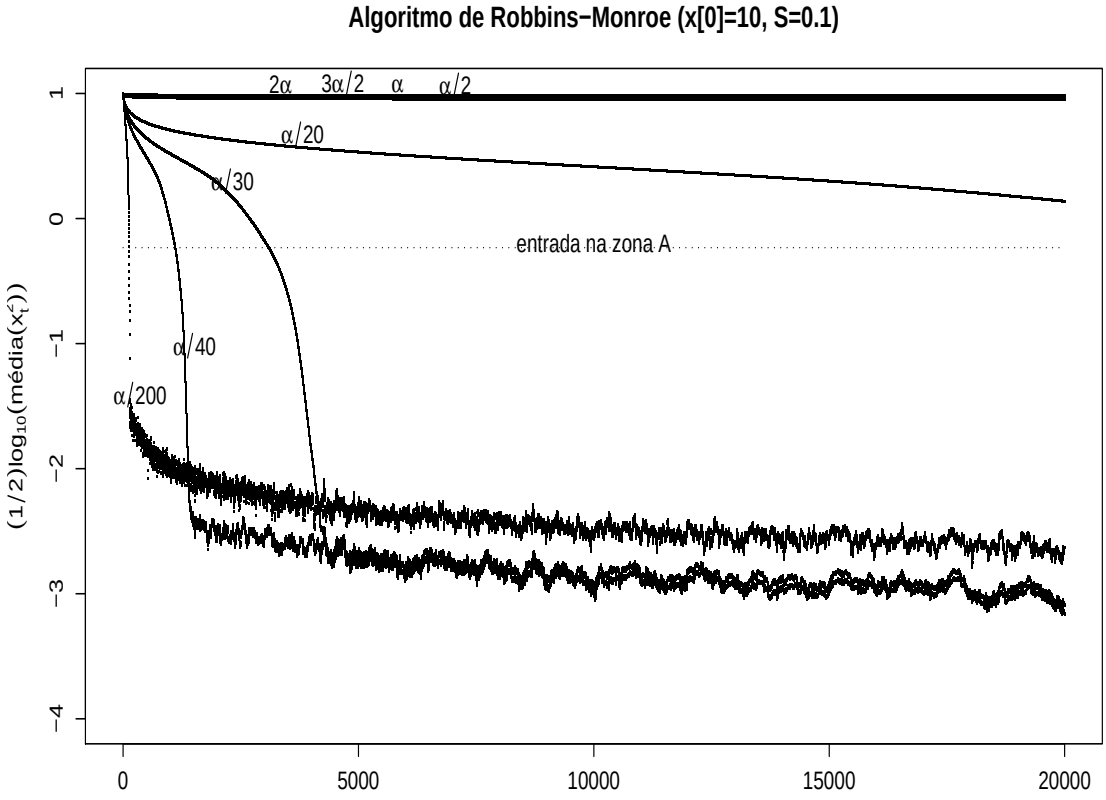


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

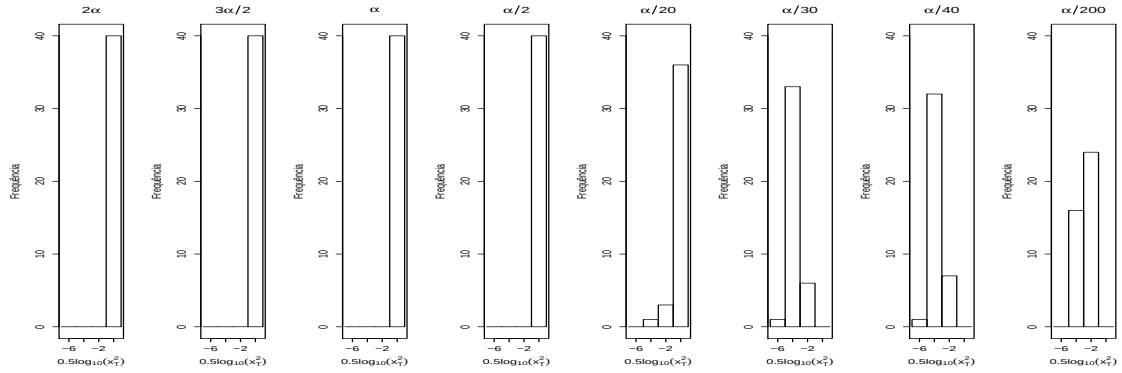


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.5: Descrição estatística da aplicação de várias configurações do algoritmo de Robbins-Monroe na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 5$ (afastado do zero da função).

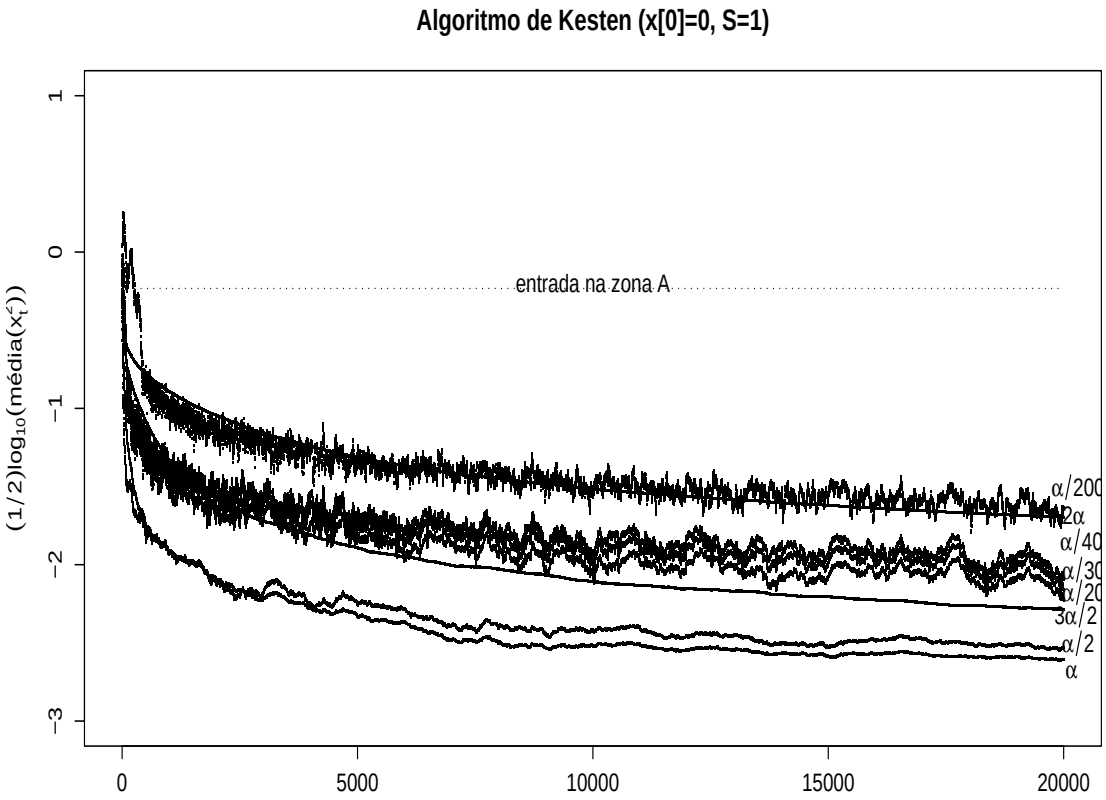


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

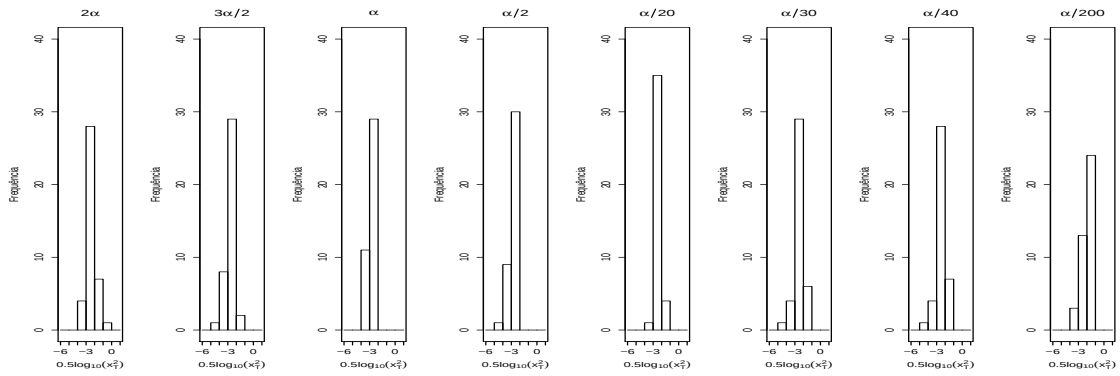


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.6: Descrição estatística da aplicação de várias configurações do algoritmo de Robbins-Monroe na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 0.1, com o processo iniciado em $x_0 = 10$ (afastado do zero da função).

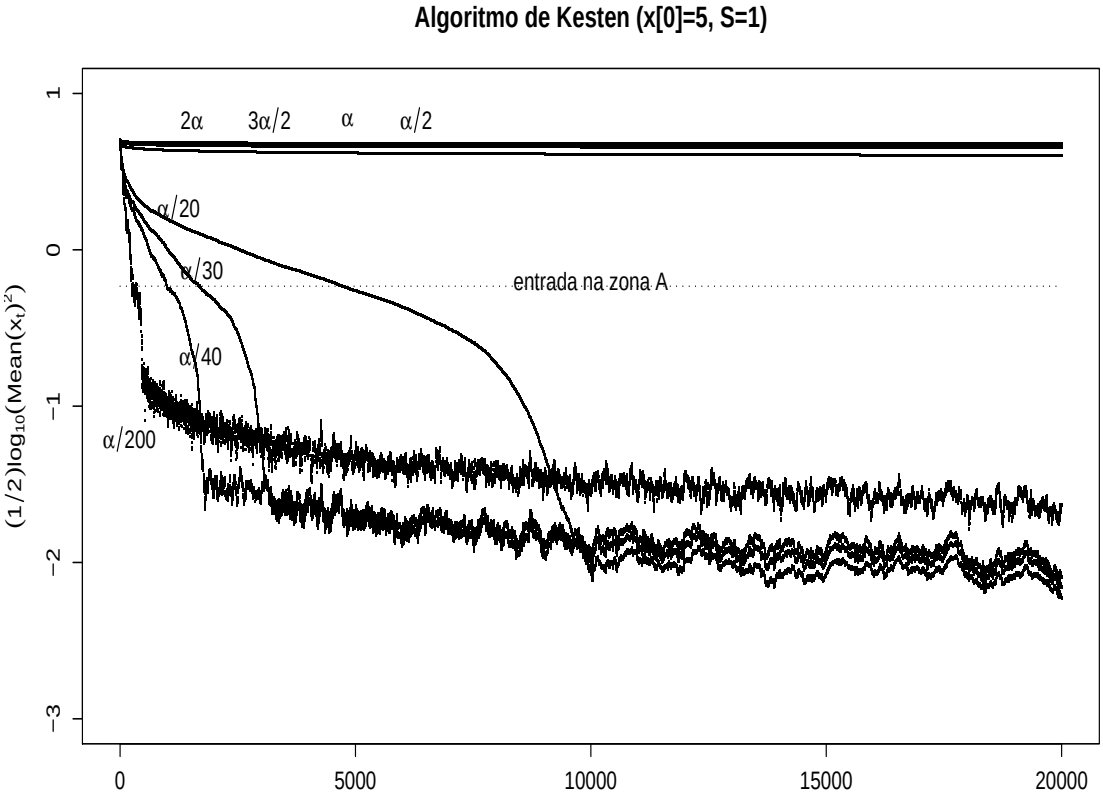


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

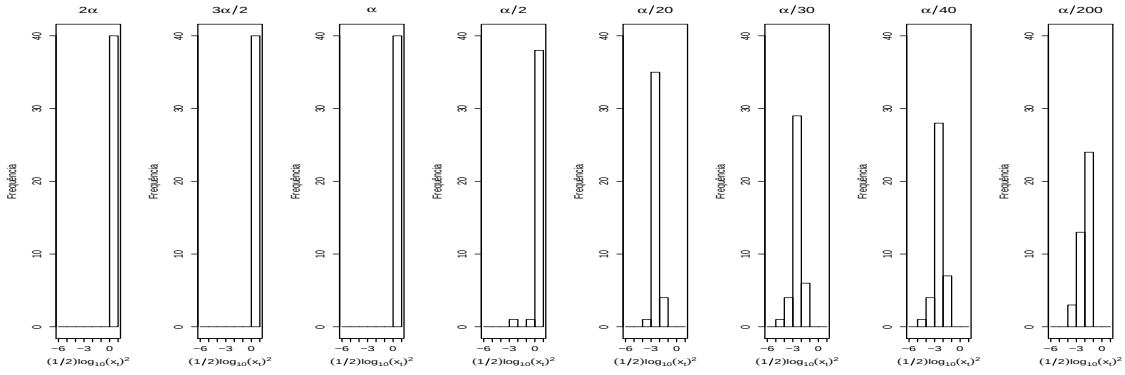


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.7: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 0$ (zero da função).

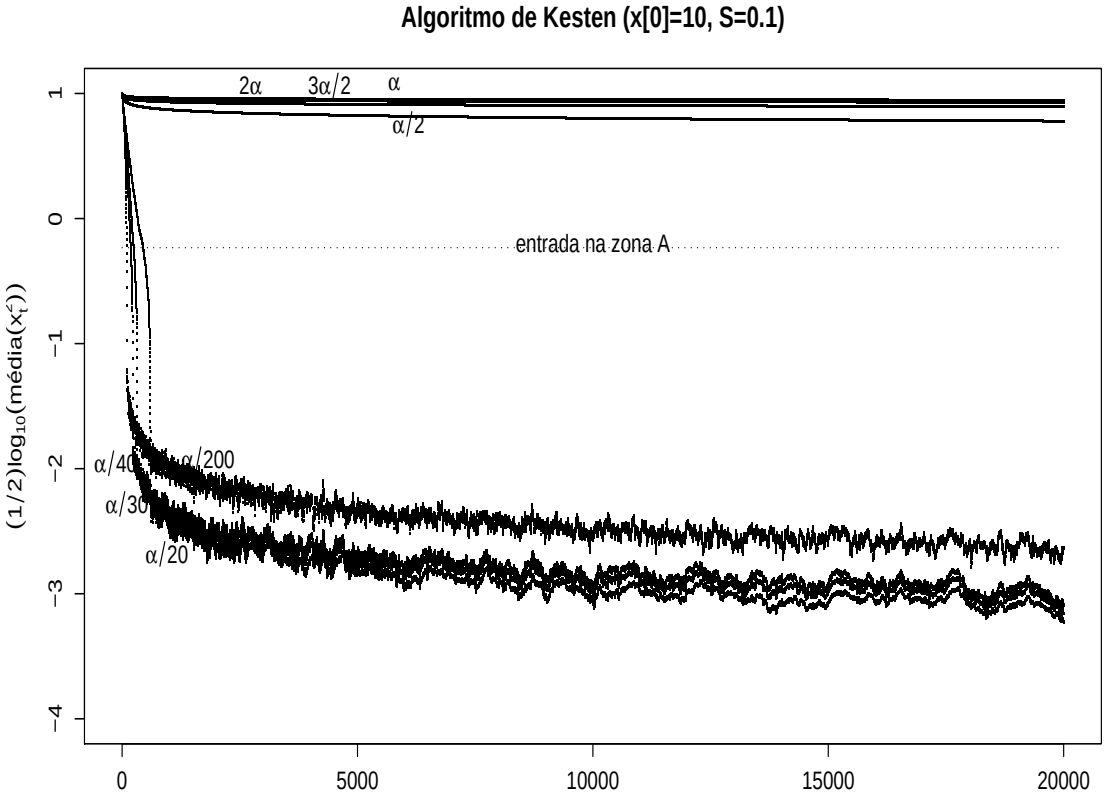


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

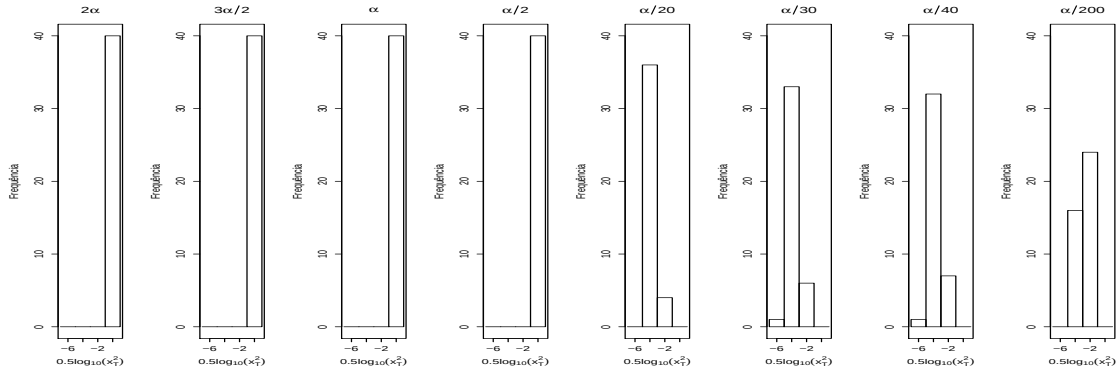


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.8: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 5$ (afastado do zero da função).

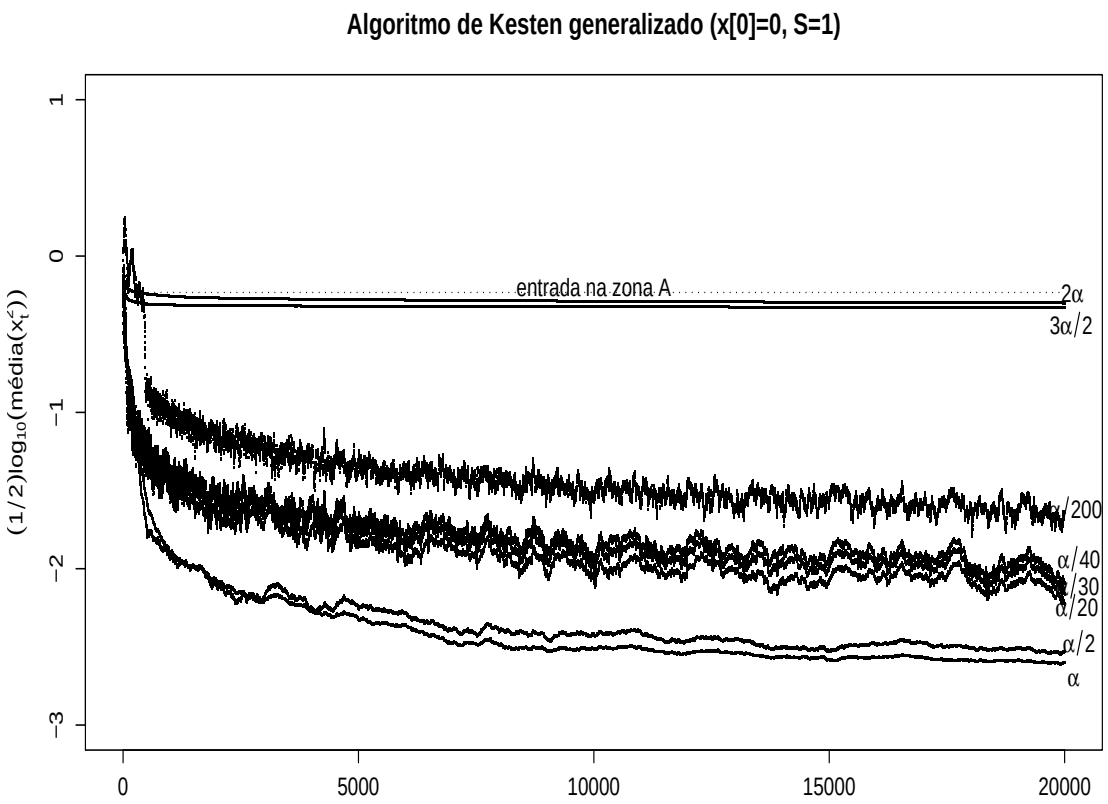


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

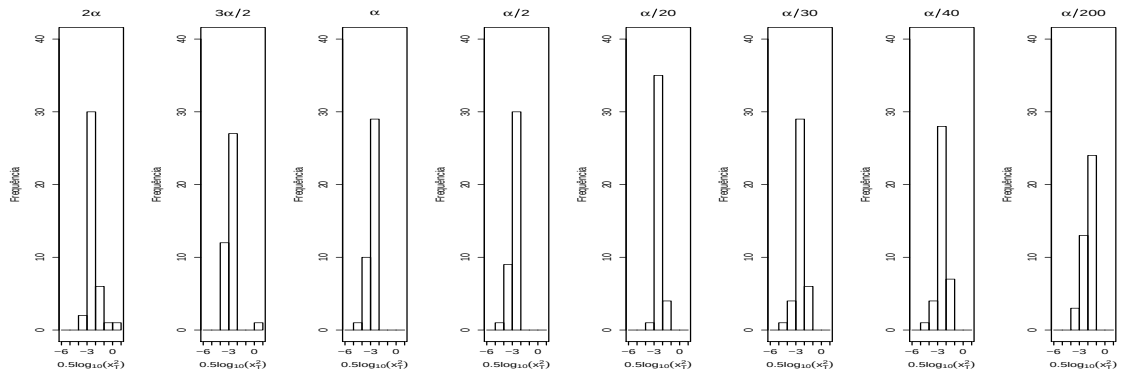


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.9: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 0.1, com o processo iniciado em $x_0 = 10$ (afastado do zero da função).

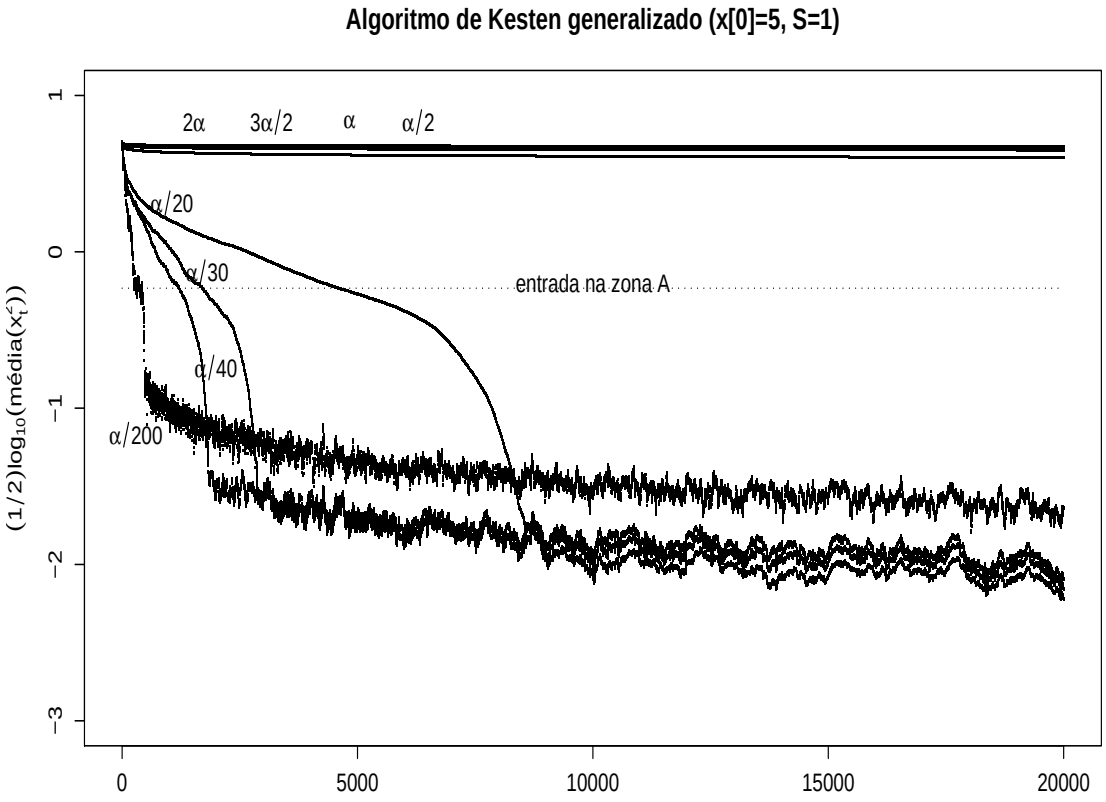


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

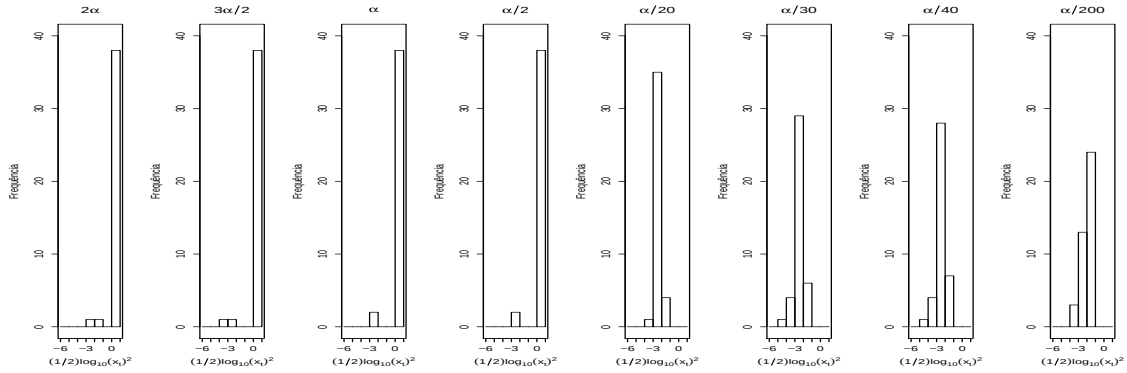


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.10: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 0$ (zero da função).

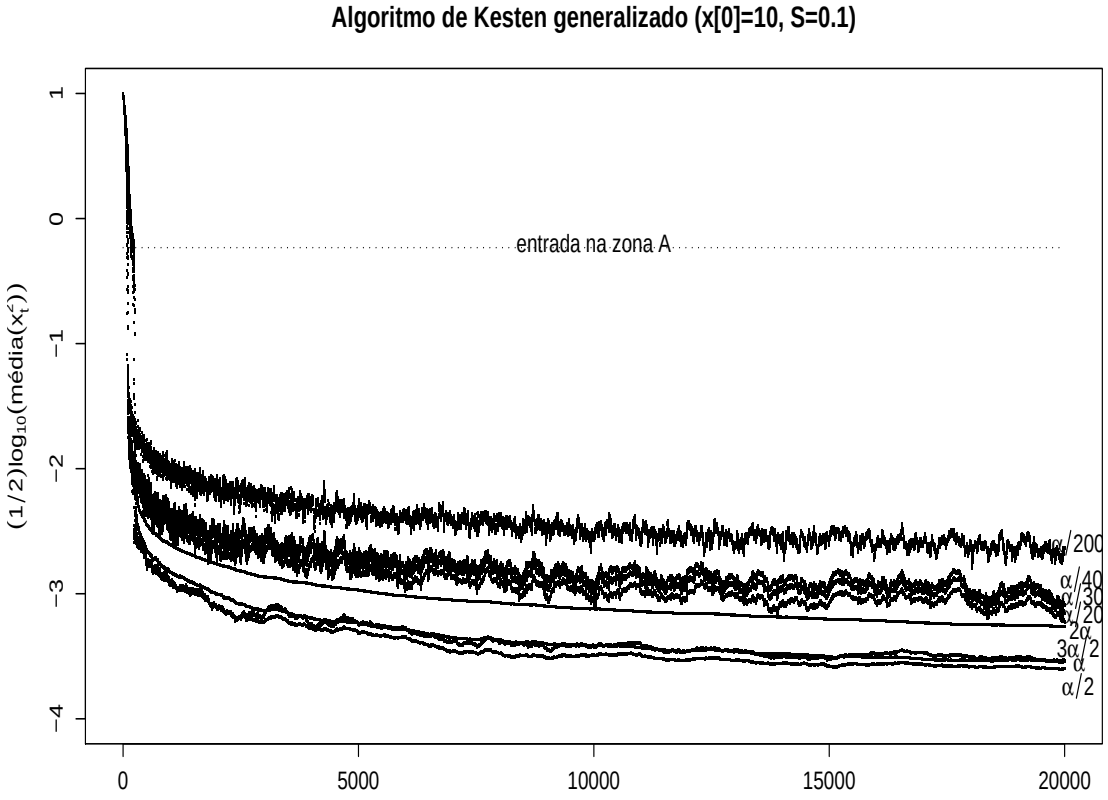


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

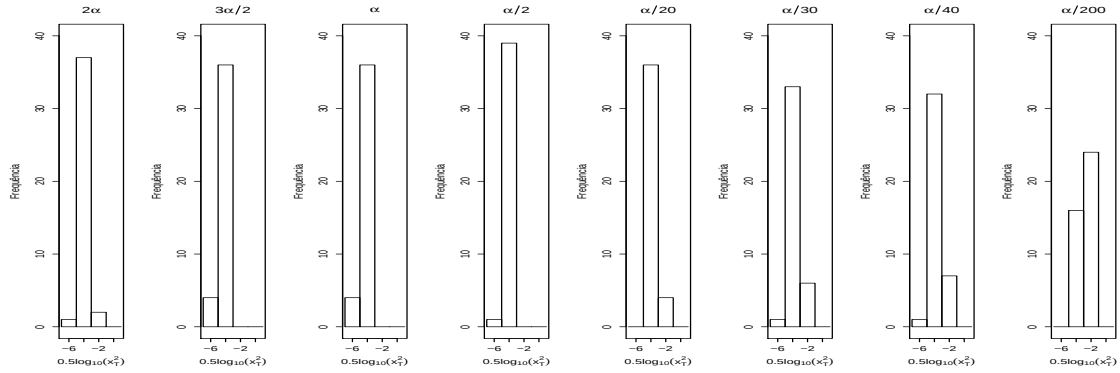


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.11: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten generalizado na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 5$ (afastado do zero da função).

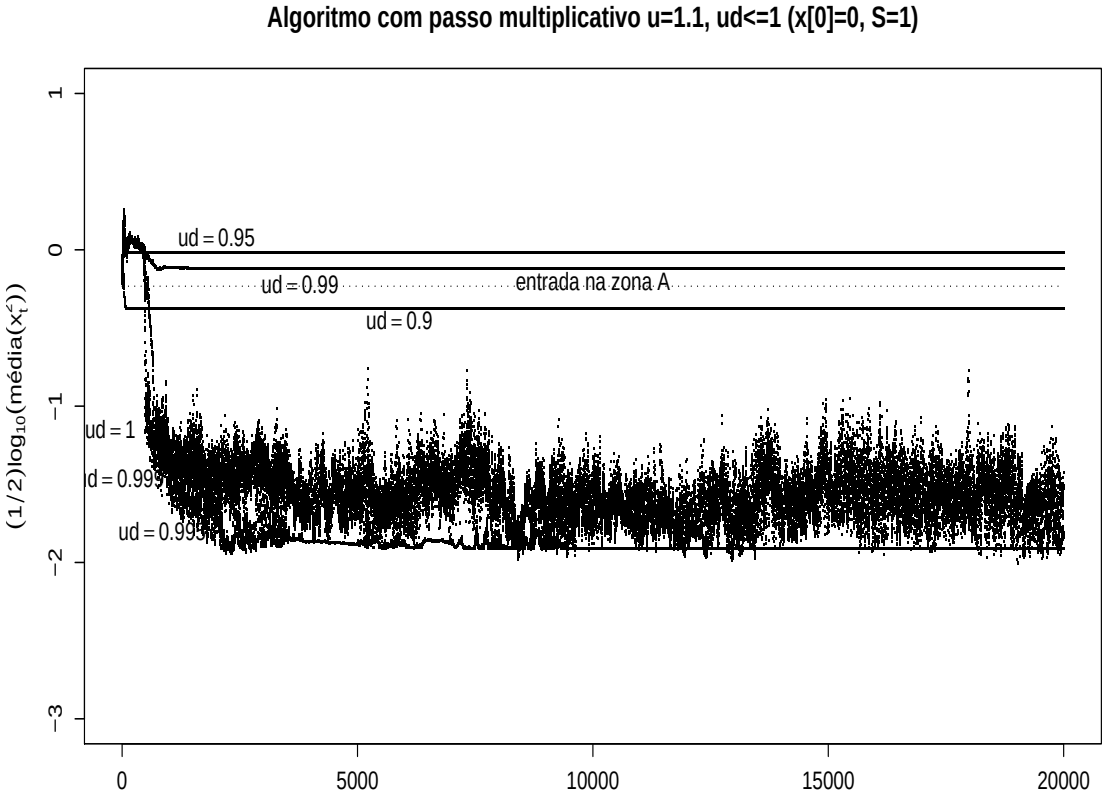


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

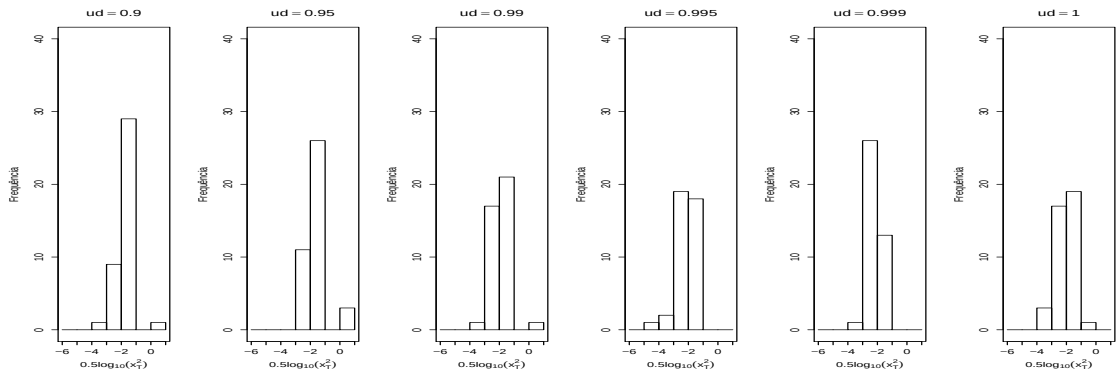


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.12: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten generalizado na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 0.1, com o processo iniciado em $x_0 = 10$ (afastado do zero da função).

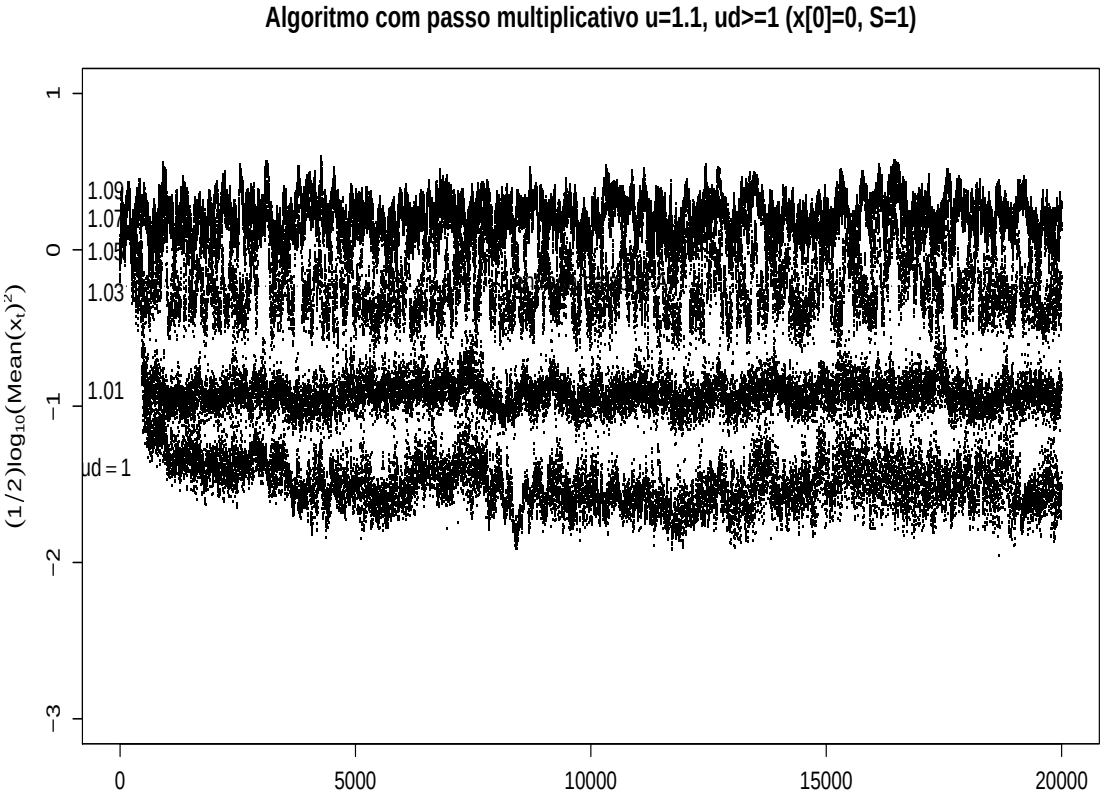


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

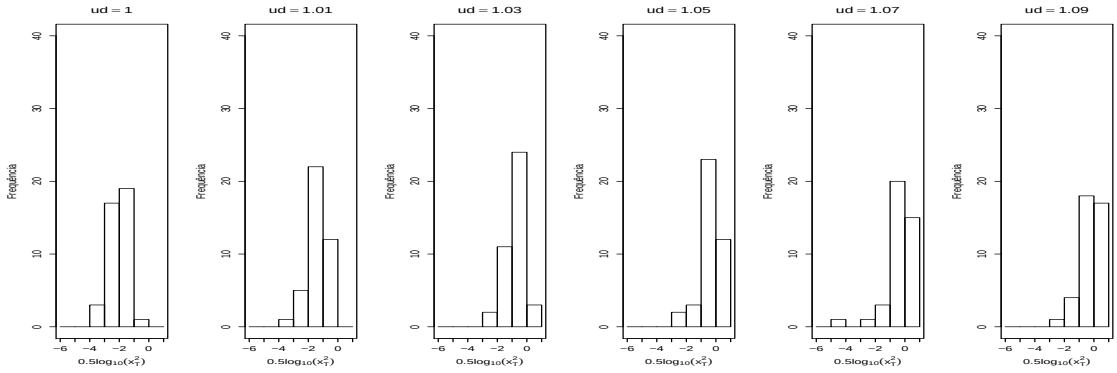


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.13: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 0.9$, $ud = 0.95$, $ud = 0.99$, $ud = 0.995$, $ud = 0.999$, $ud = 1$, com $u = 1.1$ fixo, na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 0$ (zero da função).

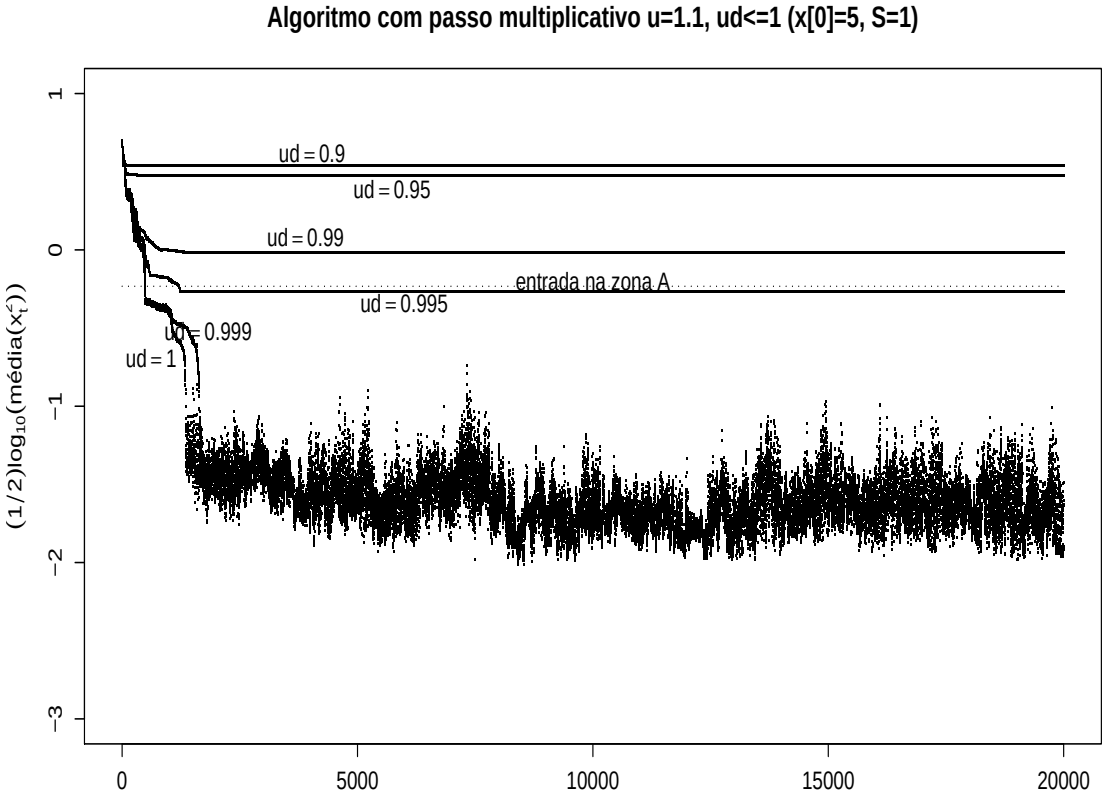


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

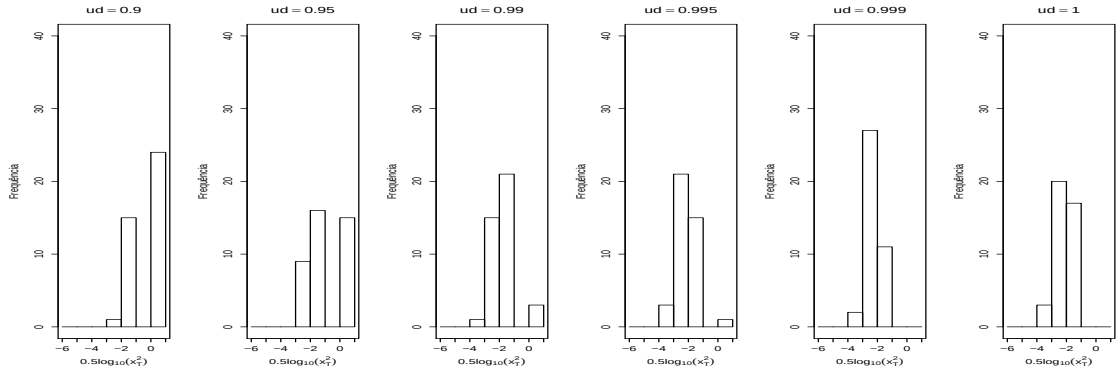


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.14: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 1$, $ud = 1.01$, $ud = 1.03$, $ud = 1.05$, $ud = 1.07$, $ud = 1.09$, com $u = 1.1$ fixo, na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 0$ (zero da função).

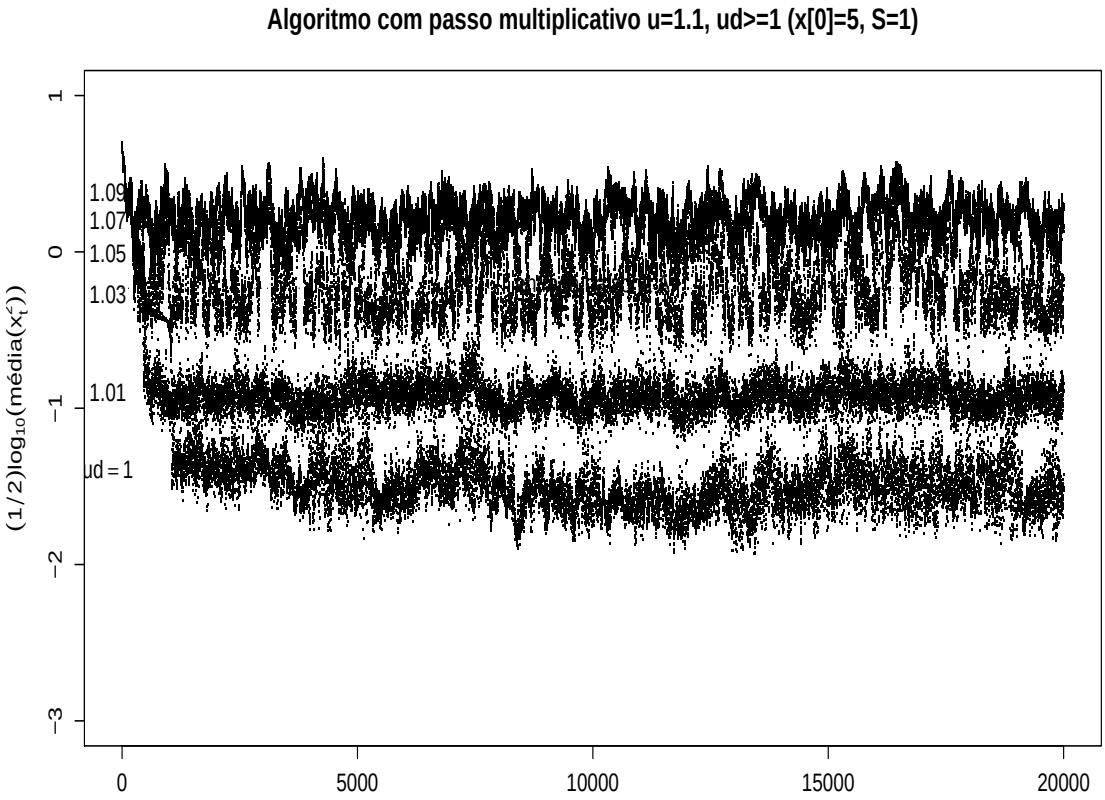


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

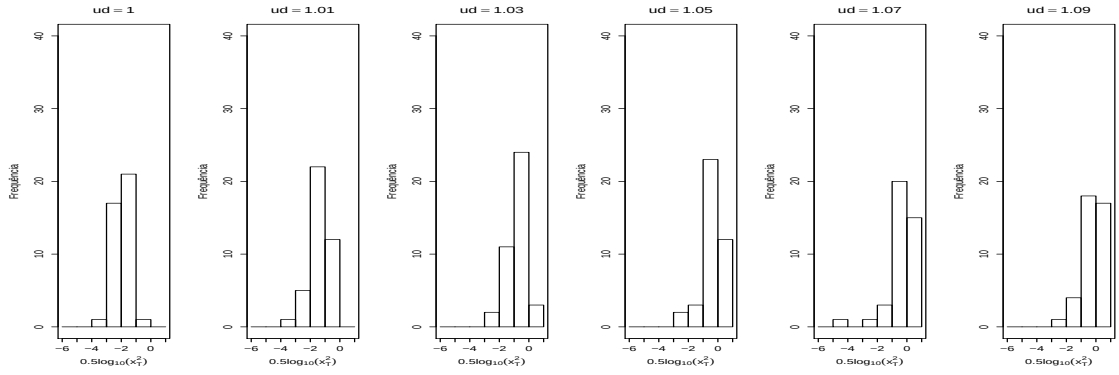


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.15: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 0.9$, $ud = 0.95$, $ud = 0.99$, $ud = 0.995$, $ud = 0.999$, $ud = 1$, com $u = 1.1$ fixo, na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 5$ (afastado do zero da função).

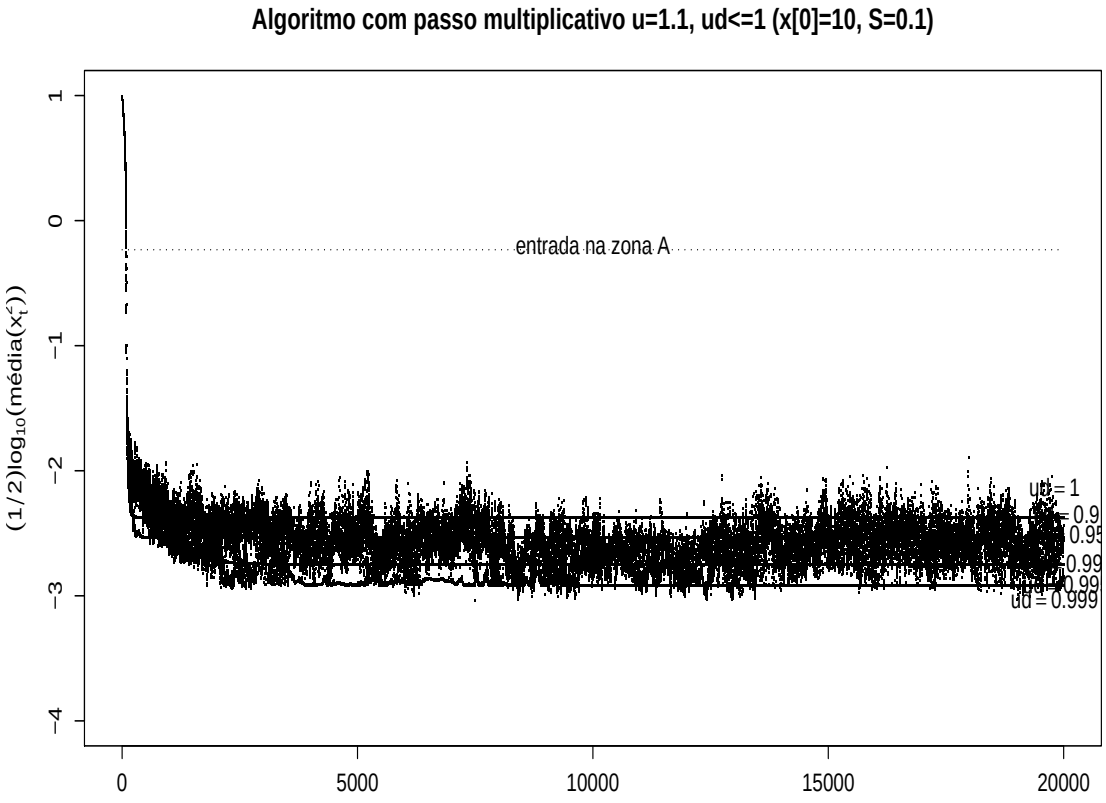


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

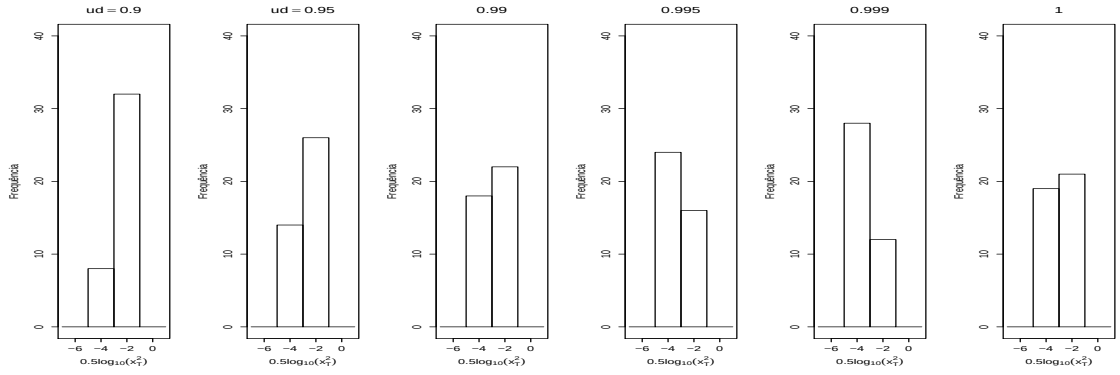


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.16: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 1$, $ud = 1.01$, $ud = 1.03$, $ud = 1.05$, $ud = 1.07$, $ud = 1.09$, com $u = 1.1$ fixo, na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = 5$ (afastado do zero da função).

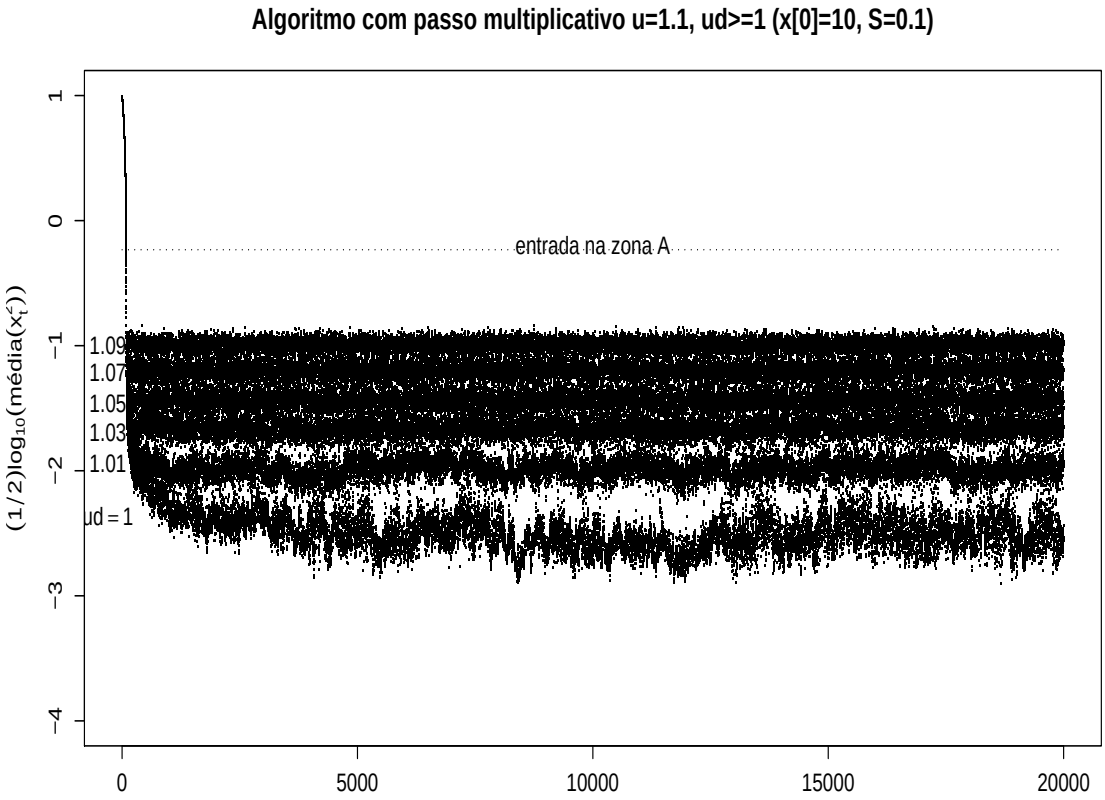


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

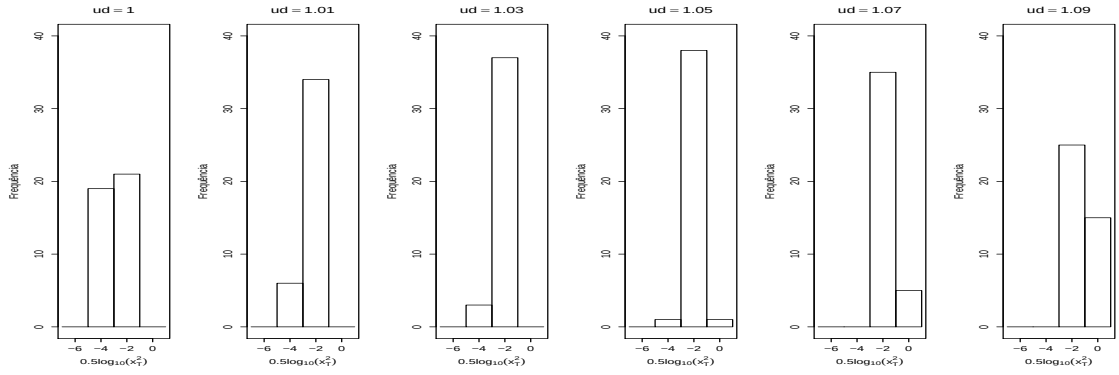


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.17: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 0.9$, $ud = 0.95$, $ud = 0.99$, $ud = 0.995$, $ud = 0.999$, $ud = 1$, com $u = 1.1$ fixo, na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 0.1, com o processo iniciado em $x_0 = 10$ (afastado do zero da função).

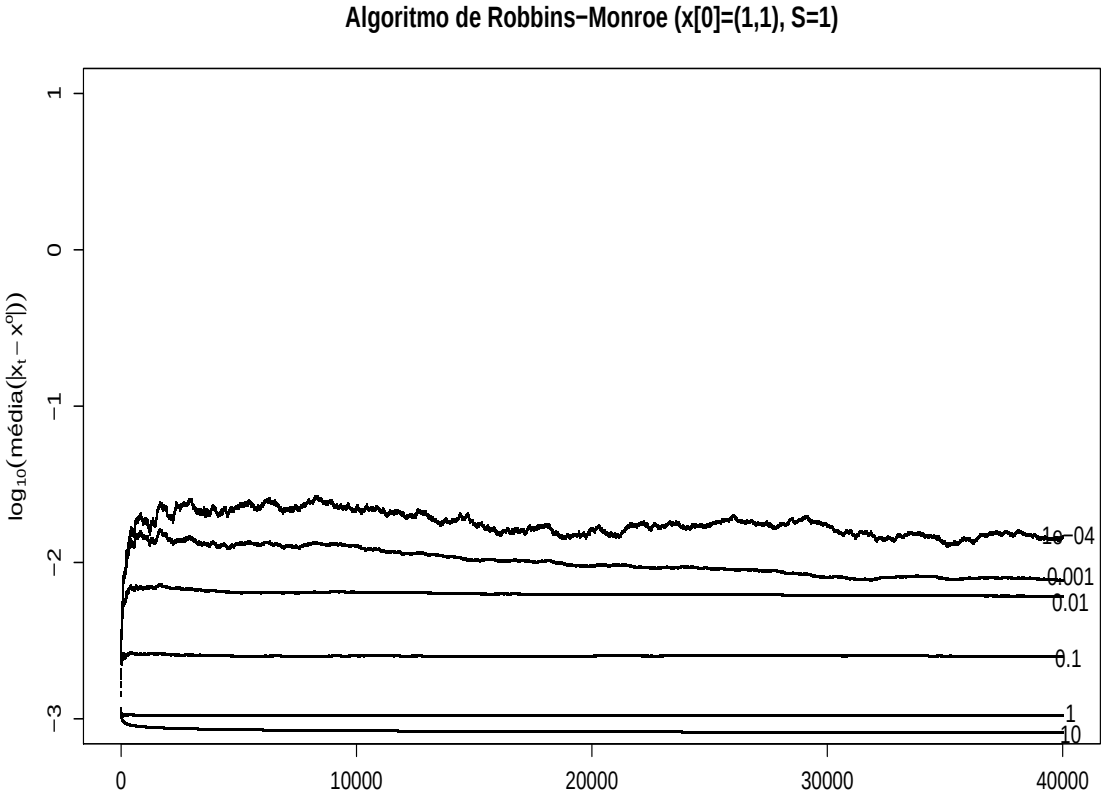


(a) Trajecto médio em 40 repetições na forma $(1/2)\log_{10}\overline{x_t^2}$ para cada configuração do algoritmo durante 20000 iterações.

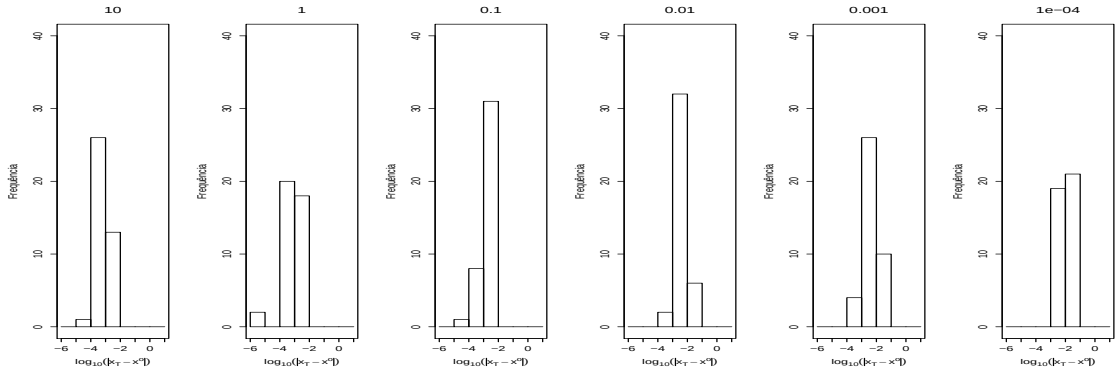


(b) Histograma da última observação na forma $(1/2)\log_{10} x_{20000}^2$.

Figura 4.18: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 1$, $ud = 1.01$, $ud = 1.03$, $ud = 1.05$, $ud = 1.07$, $ud = 1.09$, com $u = 1.1$ fixo, na determinação do zero da função $\text{sen}(\alpha \tanh(x))$, com $\alpha = 19/20\pi$, quando a medição é sujeita a um erro normal com desvio 0.1, com o processo iniciado em $x_0 = 10$ (afastado do zero da função).

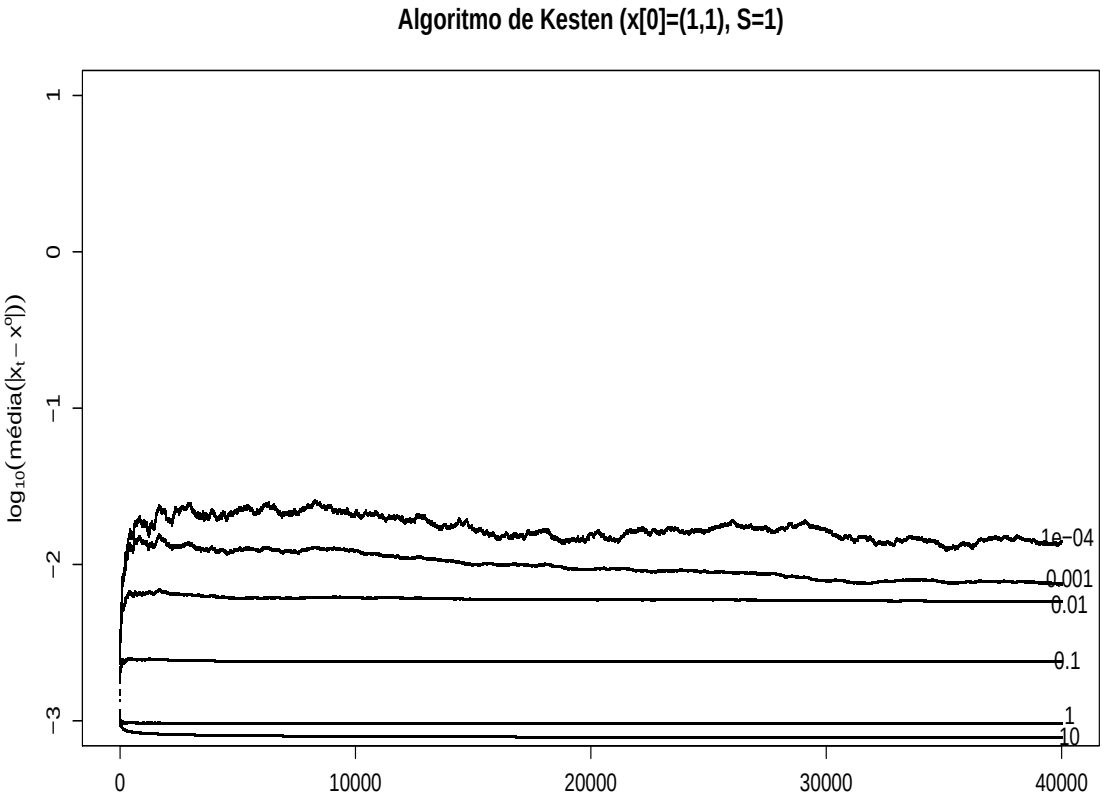


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

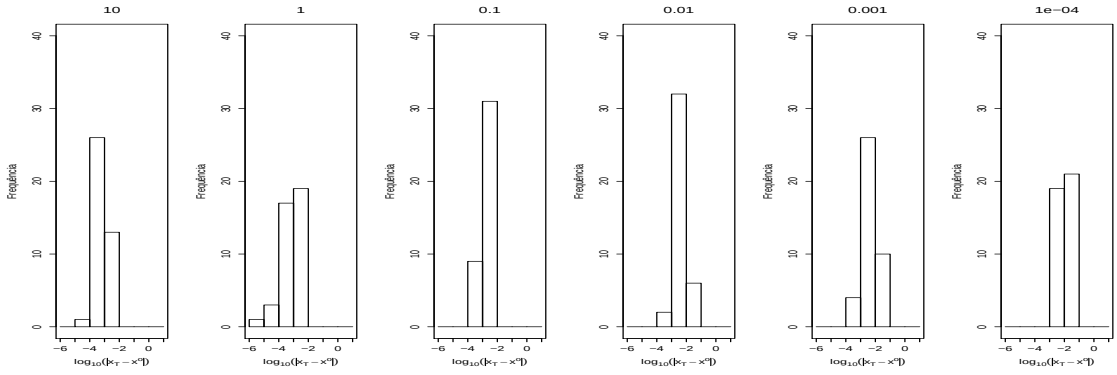


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.19: Descrição estatística da aplicação de várias configurações do algoritmo de Robbins-Monroe na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (1,1)$ (zero do gradiente).

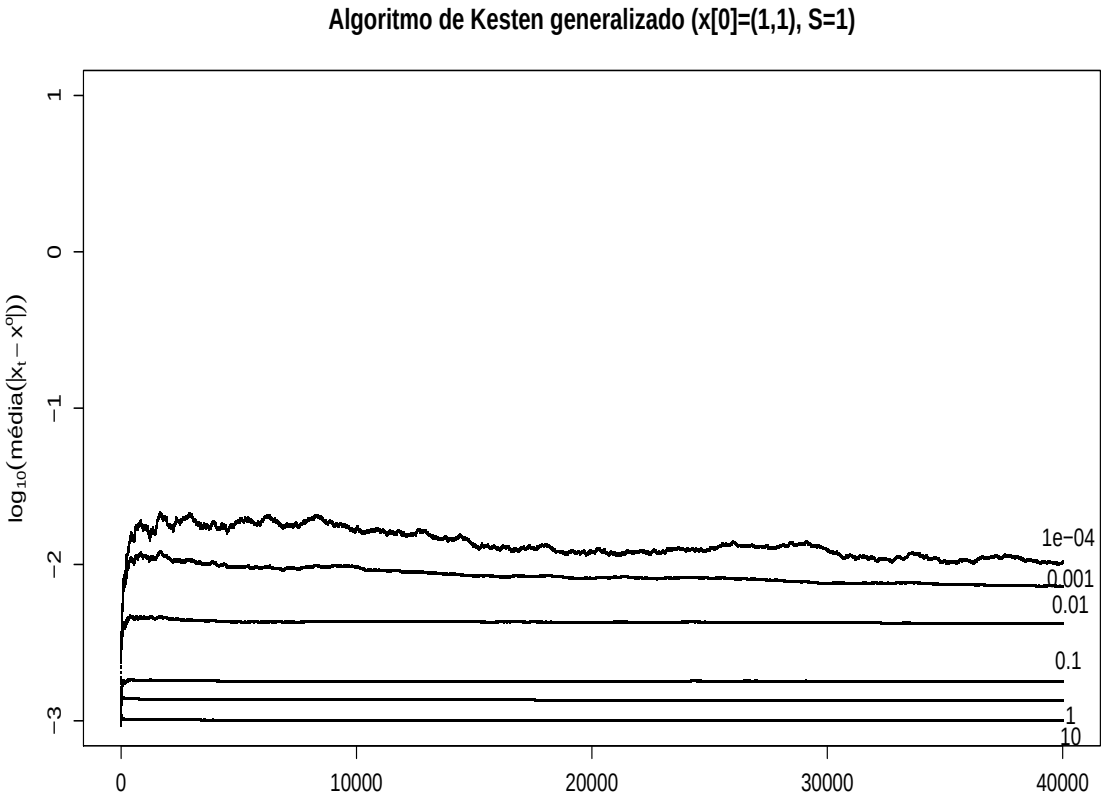


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

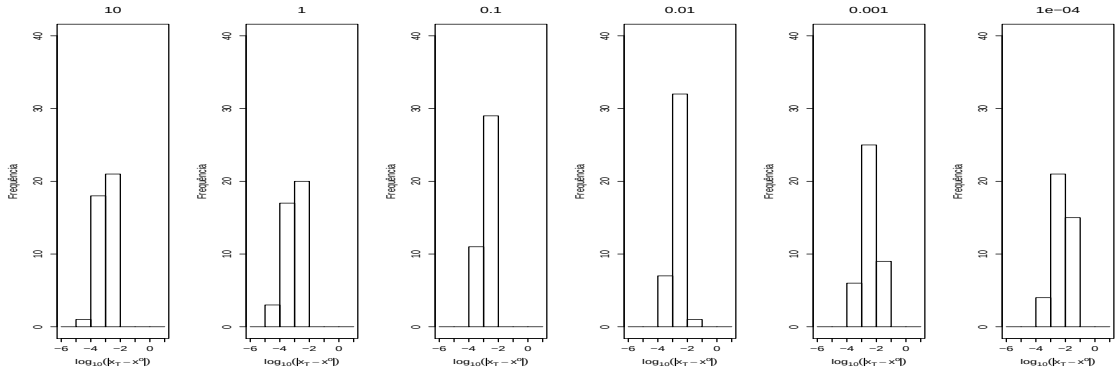


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.20: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (1,1)$ (zero do gradiente).

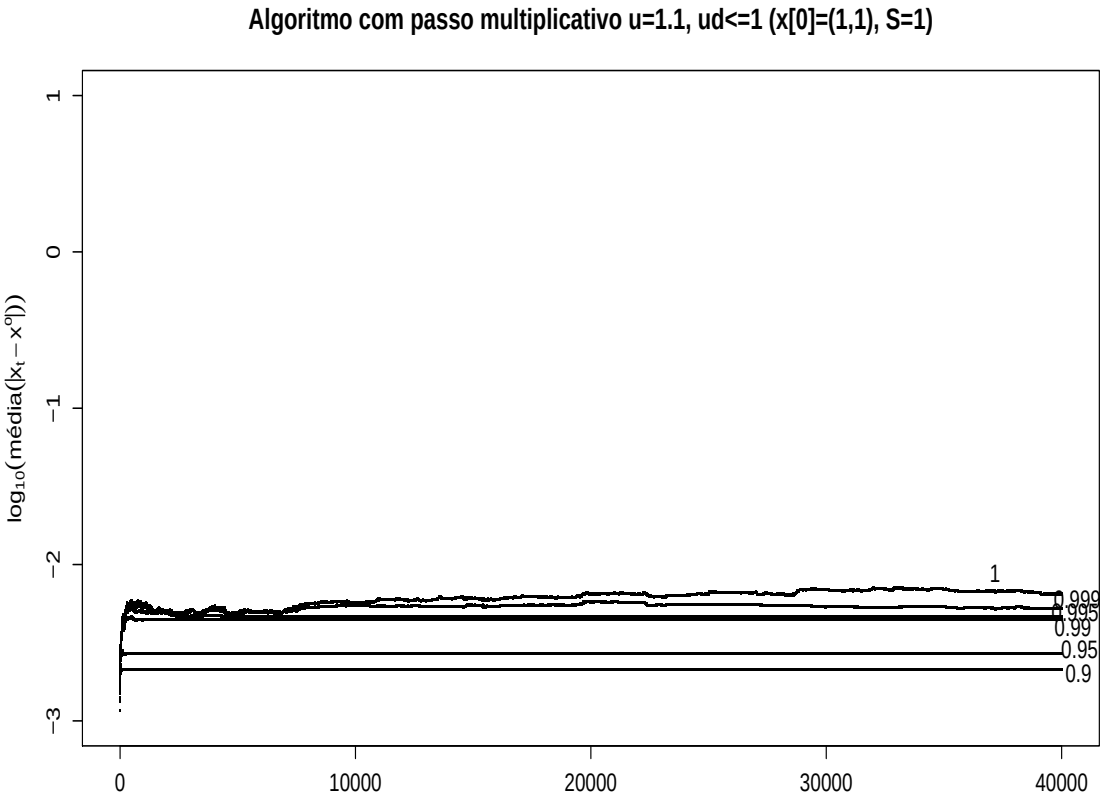


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

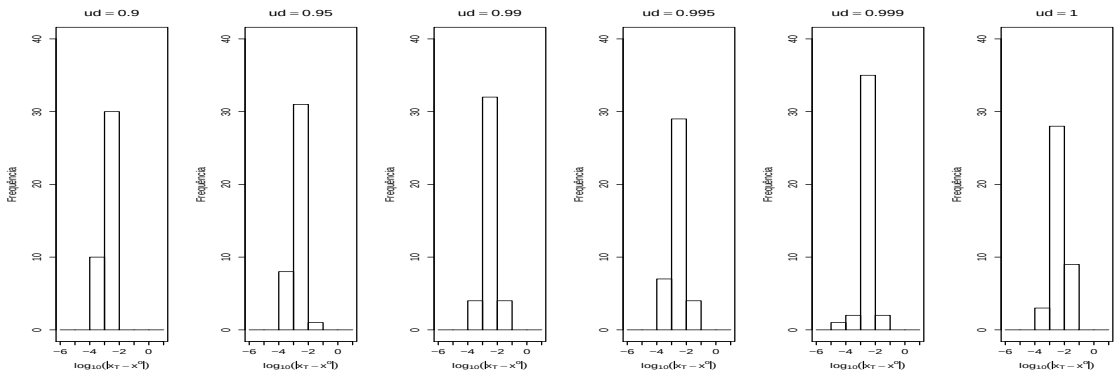


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.21: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten generalizado na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (1,1)$ (zero do gradiente).

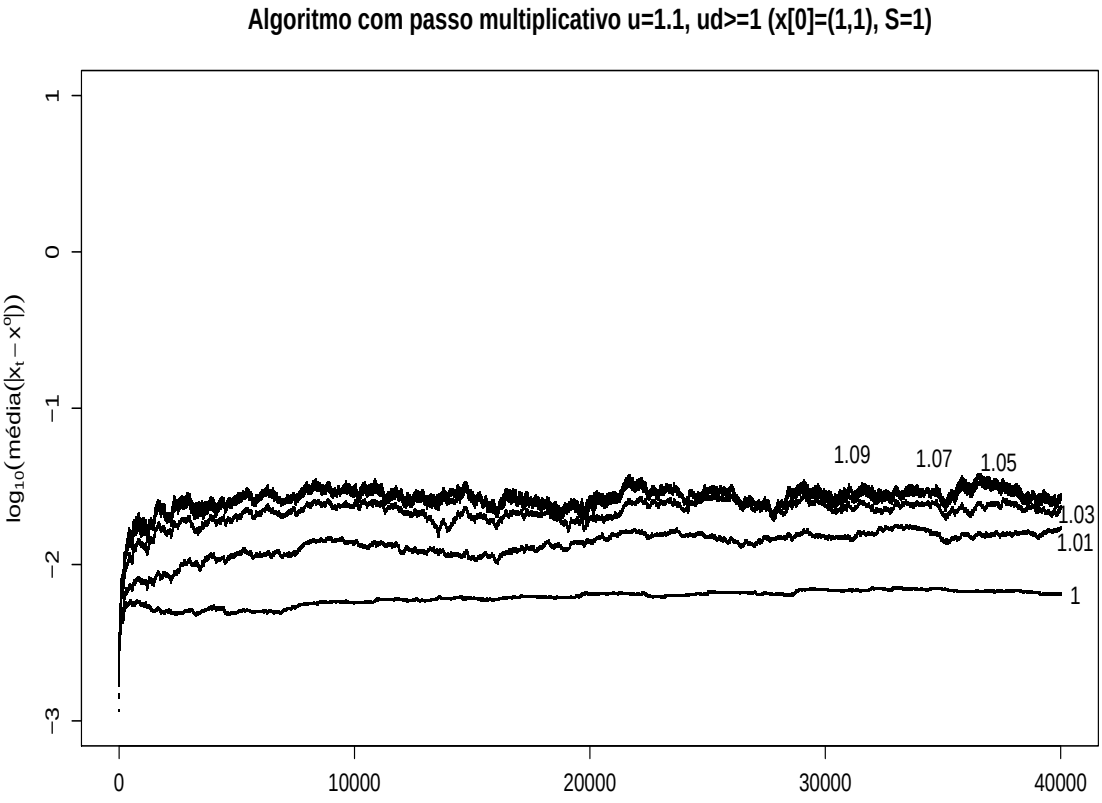


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

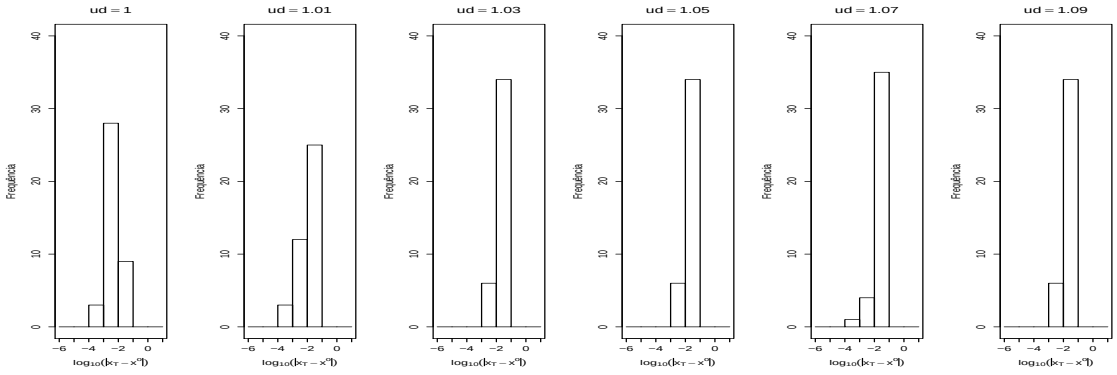


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.22: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 0.9$, $ud = 0.95$, $ud = 0.99$, $ud = 0.995$, $ud = 0.999$, $ud = 1$, com $u = 1.1$ fixo, na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (1,1)$ (zero da gradiente).

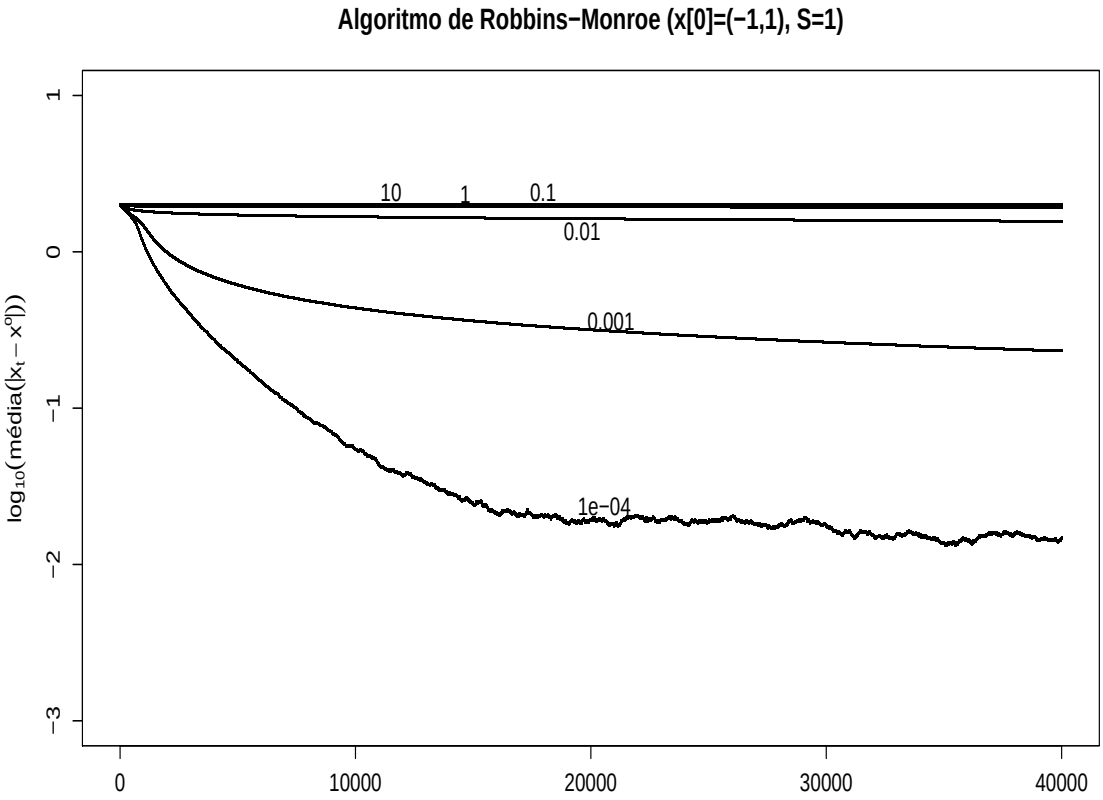


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

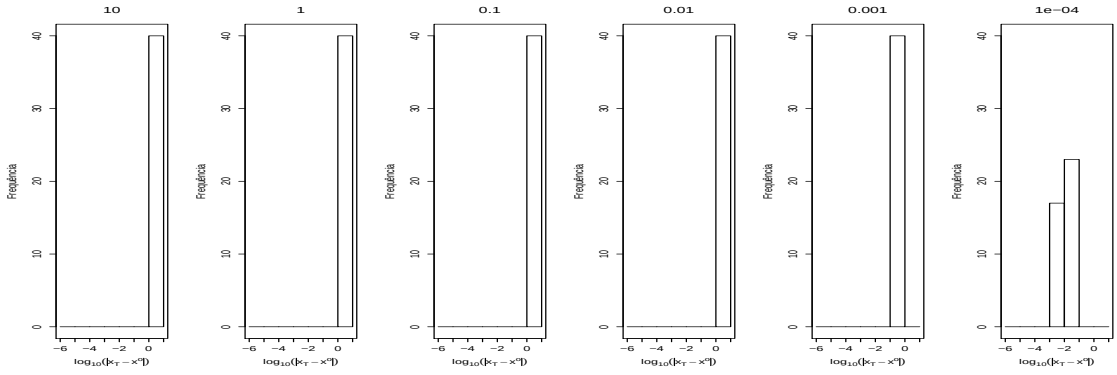


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.23: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 1$, $ud = 1.01$, $ud = 1.03$, $ud = 1.05$, $ud = 1.07$, $ud = 1.09$, com $u = 1.1$ fixo, na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (1,1)$ (zero da gradiente).

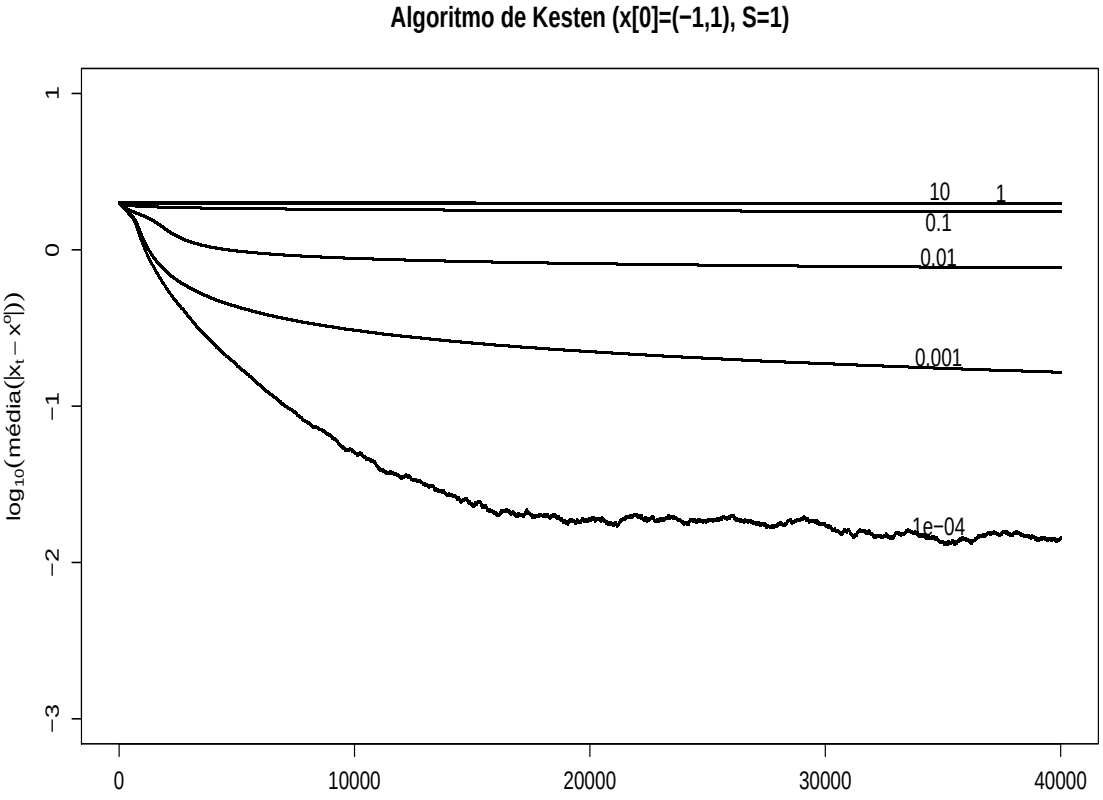


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

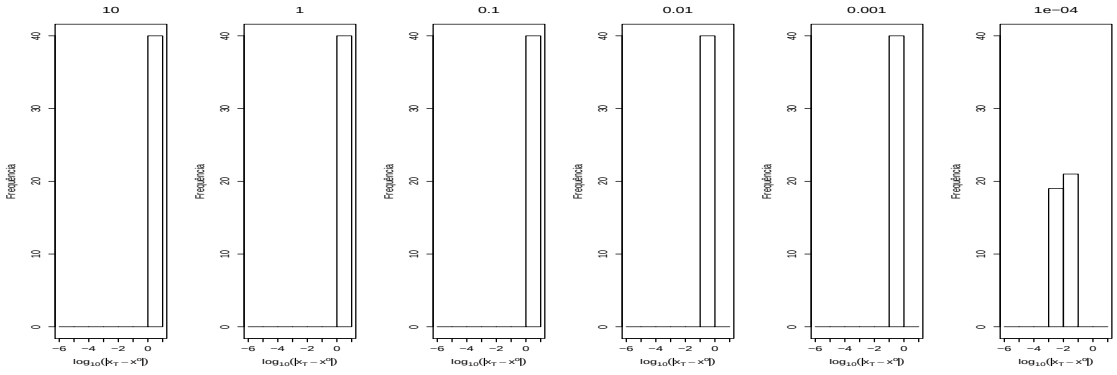


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.24: Descrição estatística da aplicação de várias configurações do algoritmo de Robbins-Monroe na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (-1, 1)$ (afastado do zero do gradiente).

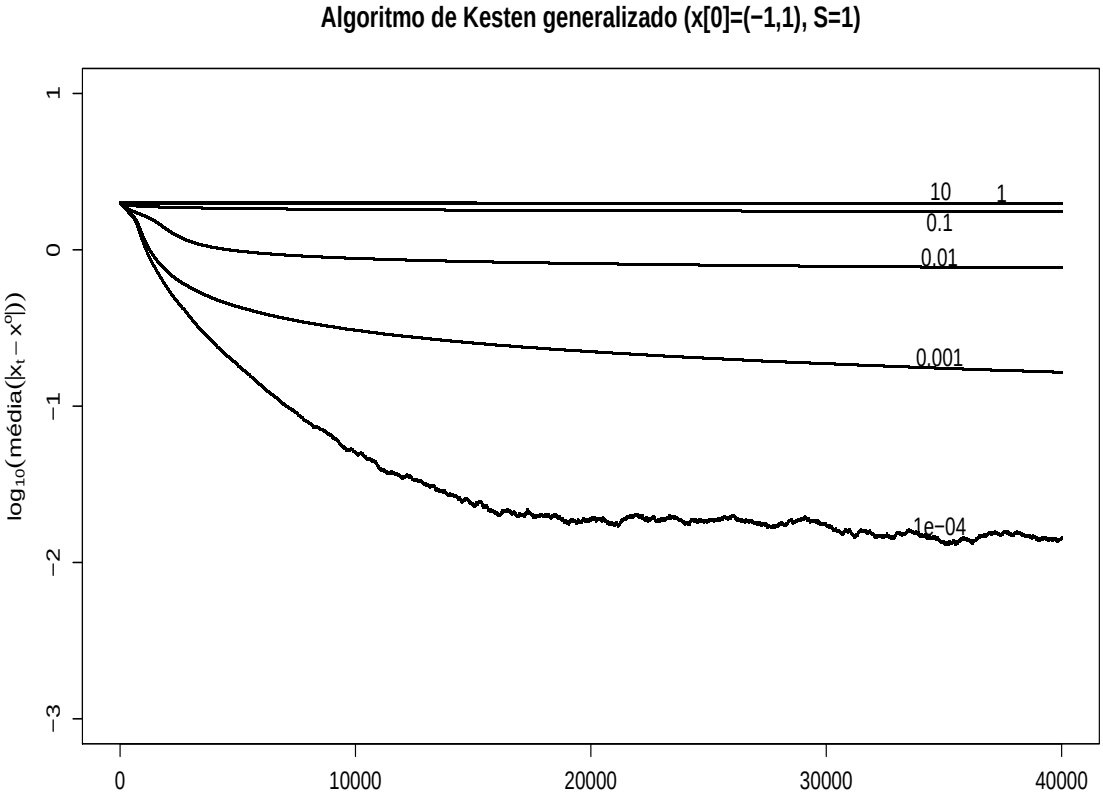


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

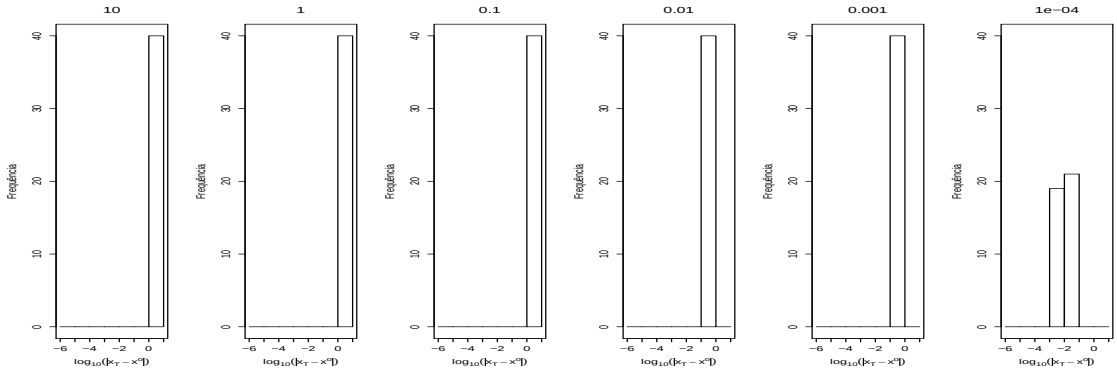


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.25: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (-1, 1)$ (afastado do zero do gradiente).

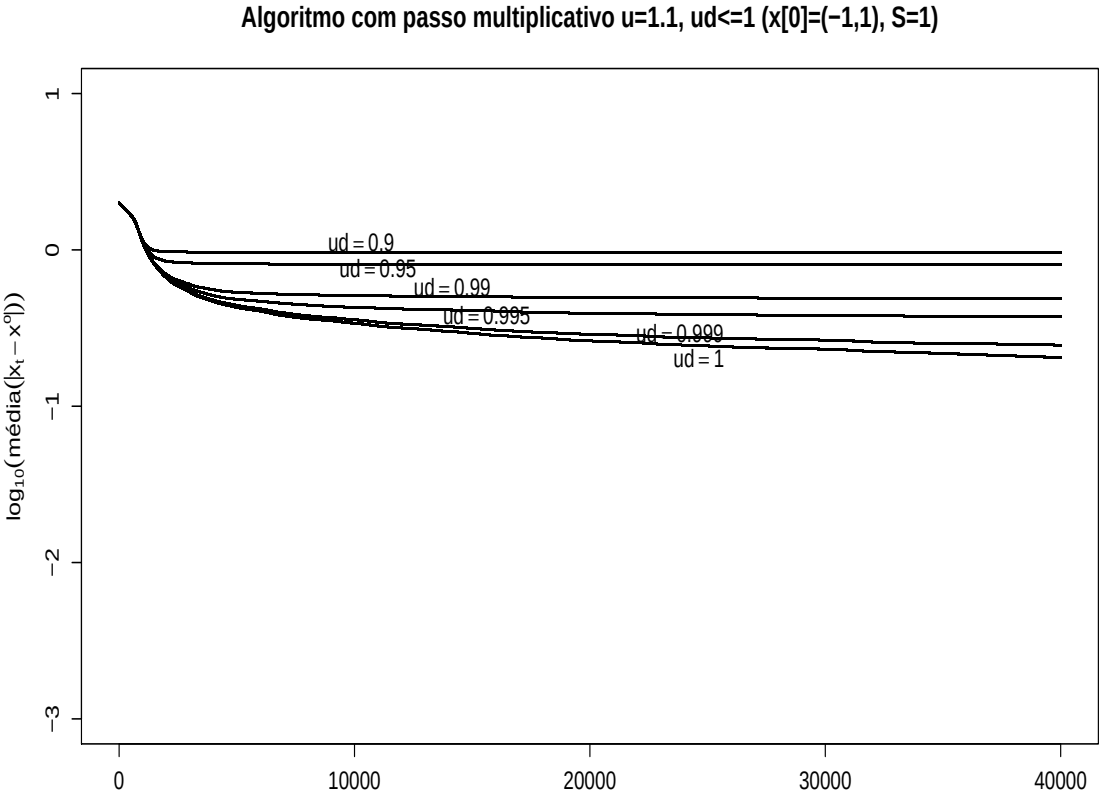


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

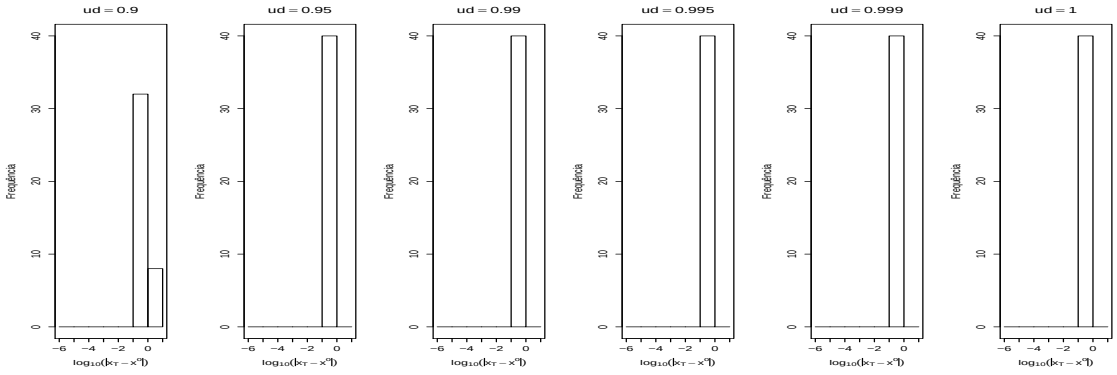


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.26: Descrição estatística da aplicação de várias configurações do algoritmo de Kesten generalizado na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (-1, 1)$ (afastado do zero do gradiente).

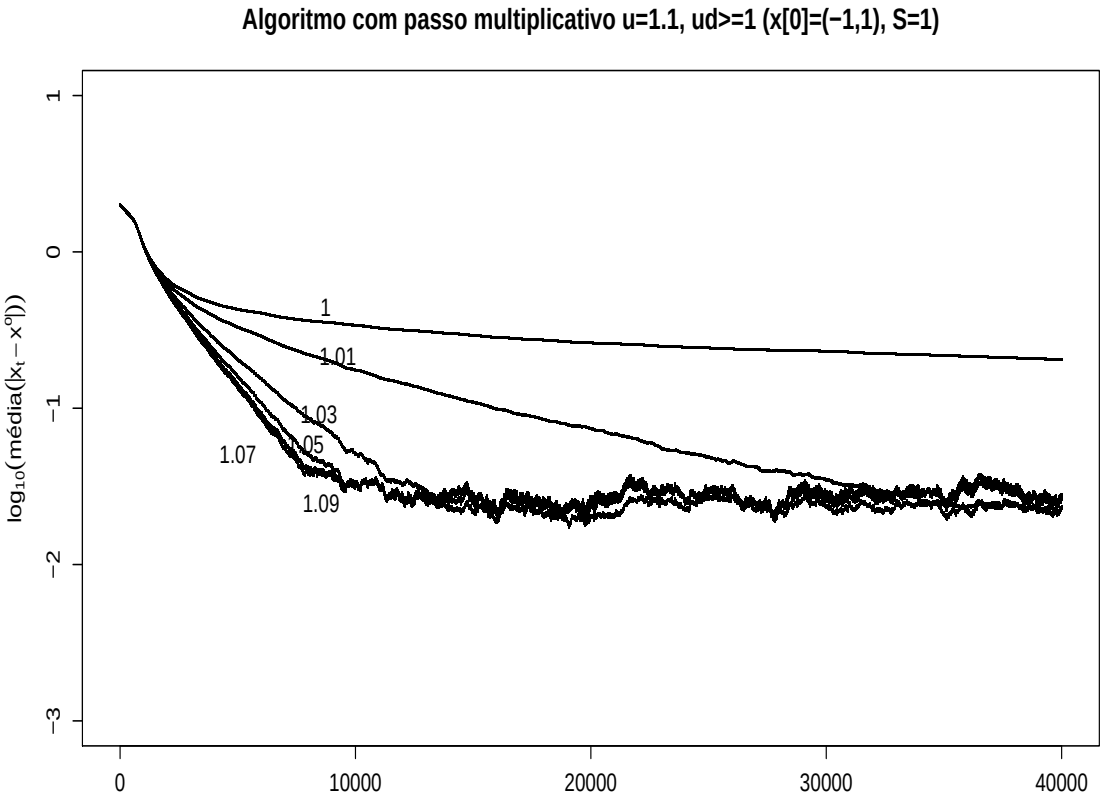


(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.

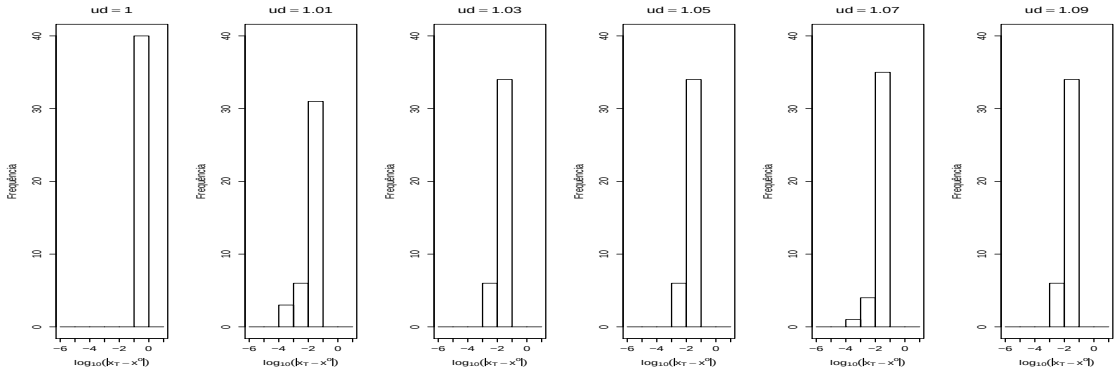


(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.27: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 0.9$, $ud = 0.95$, $ud = 0.99$, $ud = 0.995$, $ud = 0.999$, $ud = 1$, com $u = 1.1$ fixo, na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (-1, 1)$ (afastado do zero da gradiente).



(a) Trajecto médio em 40 repetições na forma $\log_{10} \overline{|x_t - (1,1)|}$ para cada configuração do algoritmo durante 40000 iterações.



(b) Histograma da última observação na forma $\log_{10} |x_{40000} - (1,1)|$.

Figura 4.28: Descrição estatística da aplicação do algoritmo de passo multiplicativo com as configurações $ud = 1$, $ud = 1.01$, $ud = 1.03$, $ud = 1.05$, $ud = 1.07$, $ud = 1.09$, com $u = 1.1$ fixo, na determinação do zero do gradiente da função de Rosenbrock (4.2), quando a medição de cada coordenada é sujeita a um erro normal com desvio 1, com o processo iniciado em $x_0 = (-1, 1)$ (afastado do zero da gradiente).

Bibliografia

[1] Luís Borges Almeida, Thibauld Langlois, José D. Amaral, and Alexander Plakhov. Parameter adaptation in stochastic optimization. In David Saad, editor, *Online Learning in Neural Networks*, pages 111–134. Cambridge University Press, Cambridge, MA, 1998.

[2] Roberto Battiti. Accelerated backpropagation learning: Two optimization methods. *Complex Systems*, 3:331–342, 1989.

[3] Roberto Battiti. First and second order methods for learning: between steepest descent and newton's method. *Neural Computation*, 4(2):141–166, 1992.

[4] Albert Benveniste, Michel Metivier, and Pierre Priouret. *Algorithmes Adaptifs et Approximations Stochastiques. Théorie et applications à l'identification, au traitement du signal et à la reconnaissance des formes. (Adaptive algorithms and stochastic approximations. Theory and applications to identification, signal processing and pattern recognition).* Techniques Stochastiques. Paris.: Masson. XIII, 367 p.;, 1987.

[5] Yvonne M.M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete multivariate analysis: Theory and practice. With the collaboration of Richard J. Light and Frederick Mosteller.* Cambridge, Mass. - London: The MIT Press. X, 557 p., 1975.

[6] Julius R. Blum. Approximation methods which converge with probability one. *Ann. Math. Stat.*, 25:382–386, 1954.

[7] Julius R. Blum. Multidimensional stochastic approximation methods. *Ann. Math. Stat.*, 25:737–744, 1954.

[8] Richard L. Burden and J.Douglas Faires. *Numerical Analysis. 5th ed.* Boston, MA: PWS Publishing Company. London: ITP International Thomson Publishing, XIV, 768 p., 1993.

[9] Tai-Cong Chen. Acceleration of Levenberg-Marquardt training of neural networks with variable decay rate. In *Neural Networks, 2003. Proceedings of the International Joint Conference on, 20-24 July*, volume 3, pages 1873–1878, Massachusetts – Washington, 2003. IEEE.

[10] K.L. Chung. On a stochastic approximation method. *Ann. Math. Stat.*, 25:463–483, 1954.

[11] James Davidson. *Stochastic Limit Theory. An Introduction for Econometricians. Repr.* Advanced Texts in Econometrics. Oxford: Oxford Univ. Press. XXII, 539 p., 1997.

[12] Bernard Delyon. Stochastic approximation with decreasing gain: Convergence and asymptotic theory. Unpublished Lecture Notes, 2000.

[13] Bernard Delyon and Anatoli Juditsky. Stochastic optimization with averaging of trajectories. *Stochastics Stochastics Rep.*, 39(2-3):107–118, 1992.

[14] Bernard Delyon and Anatoli Juditsky. Accelerated stochastic approximation. *SIAM J. Optim.*, 3(4):868–881, 1993.

[15] Bernard Delyon and Anatoli Juditsky. Asymptotical study of parameter tracking algorithms. *SIAM J. Control Optimization*, 33(1):323–345, 1995.

[16] Václav Fabian. Stochastic approximation. In *Optimizing Meth. Statist., Proc. Sympos.*, pages 439–470. Ohio State Univ., 1971.

[17] Yan Fang and Terrence J. Sejnowski. Faster learning for dynamic recurrent backpropagation. *Neural Computation*, 2(3):270–273, 1990.

[18] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion.* Addison-wesley, Reading, Massachusetts, 1993.

[19] Lei Guo and Lennart Ljung. Performance analysis of general tracking algorithms. *IEEE Trans. Autom. Control*, 40(8):1388–1402, 1995.

[20] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

[21] John Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the theory of neural computation.* Addison-Wesley Longman Publishing Co., Inc., 1991.

[22] Eiichi Isogai. A stochastic approximation method approximating the roots of time varying regression functions. *Sci. Rep. Niigata Univ., Ser. A*, 21:1–18, 1985.

[23] Harry Kesten. Accelerated stochastic approximation. *Ann. Math. Stat.*, 29:41–59, 1958.

[24] J. Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.*, 23:462–466, 1952.

[25] Harold J. Kushner and G.George Yin. *Stochastic approximation algorithms and applications.* Applications of Mathematics. 35. Berlin: Springer. XXI, 417 p., 1997.

[26] Peter Lancaster and Miron Tismenetsky. *The theory of matrices. 2nd ed., with applications.* Computer Science and Applied Mathematics. Orlando etc.: Academic Press (Harcourt Brace Jovanovich, Publishers). XV, 570 p., 1985.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[28] Achim Lewandowski and Peter Protzel. Approximation of time-varying functions with local regression models. In *Dorffner, Georg (ed.) et al., Artificial neural networks - ICANN 2001. International conference, Vienna, Austria, August 21-25, 2001. Proceedings. Berlin: Springer. Lect. Notes Comput. Sci. 2130, 237-243* . Springer, 2001.

[29] Odile Macchi and Eweda Eweda. Second-order convergence analysis of stochastic adaptive linear filtering. *IEEE Trans. Autom. Control*, 28:76–85, 1983.

[30] R.E. Mahony and R. Lozano. Generalized forgetting functions for on-line least-squares identification of time-varying systems. *Int. J. Adapt. Control Signal Process.*, 15(4):393–413, 2001.

[31] M.B. Nevel'son and R.Z. Has'minskii. *Stochastic approximation and recursive estimation. Translated from the Russian by Israel Program for Scientific Translations. Translation edited by B. Silver.* Translations of Mathematical Monographs. Vol. 47. Providence, R.I.: American Mathematical Society. IV, 244 p., 1976.

[32] Athanasios Papoulis. *Probability, random variables, and stochastic processes. 2nd ed.* McGraw-Hill Series in Electrical Engineering. Communications and Information Theory. New York etc.: McGraw-Hill Book Company. XV, 576 p., 1984.

[33] Alexander Plakhov and Luís Borges Almeida. Modified kesten algorithm. Technical report, Instituto Superior Técnico, Lisbon, Portugal., 2000.

[34] Alexander Plakhov and Pedro Cruz. A stochastic approximation algorithm with multiplicative step size adaptation. Work done during second author PhD Thesis. Submitted., 2004.

[35] Alexander Plakhov and Pedro Cruz. A stochastic approximation algorithm with step size adaptation. *Journal of Mathematical Sciences – Special Volume "Aveiro Seminar on Control, Optimization and Graph Theory"*, 107:119–130, 2004.

[36] B.T. Polyak. New stochastic approximation type procedures. *Autom. i Telemekh*, 7:98–107, 1990.

[37] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[38] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

[39] N.M. Roehl and C.E. Pedreira. An online learning approach: a methodology for time varying applications. *Neural Comput. Appl.*, 10(2):101–107, 2001.

[40] David Ruppert. Efficient estimators from a slowly convergent robbins-monro process. Technical report, Cornell University, Ithaca, NY, 1988. This work as been known as the stochastic approximation 'averaging algorithm' for asymptotical optimality.

[41] Paul Révész. On the rate of convergence of Kesten's 'accelerated stochastic approximation'. *Studia Sci. Math. Hungar.*, 9:453–460, 1974.

[42] Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Stat.*, 29:373–405, 1958.

[43] R. Salomon and van J. L. Hemmen. Accelerating backpropagation through dynamic self-adaptation. *Neural Networks*, 9(4):589–601, 1996.

[44] Fernando M. Silva and Luís B. Almeida. Speeding up backpropagation. In R. Eckmiller, editor, *Advanced Neural Computers*, pages 151–158, Amsterdam, 1990. Elsevier Science Publishers.

[45] James C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Autom. Control*, 45(10):1839–1853, 2000.

[46] Nguyen Huu Tien. On the accelerated stochastic approximation. *Stud. Sci. Math. Hung.*, 12:371–380, 1977.

[47] J.H. Venter. On Dvoretzky stochastic approximation theorems. *Ann. Math. Stat.*, 37:1534–1544, 1966.

[48] David Williams. *Probability with martingales.* Cambridge University Press. XV, 251 p., 1991.